*Journal of Mathematical Sociology*
1994, Vol. 19(4), pp. 221–262
Reprints available directly from the publisher
Photocopying permitted by license only



# THE NATURE OF THE SOCIAL AGENT*

KATHLEEN CARLEY

*Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

ALLEN NEWELL

*School of Computer Science and Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

*July 26, 1993; revised March 28, 1994*

We pose the question, *What is necessary to build an artificial social agent?* Current theories of cognition provide an analytical tool for peeling away what is understood about individual cognition so as to reveal wherein lies the social. We fractionate a set of agent characteristics to describe a Model Social Agent. The fractionation matrix is, itself, a set of increasingly inclusive models, each one a more adequate description of the social agent required by the social sciences. The fractionation reflects limits to the agent's information-processing capabilities and enrichment of the mental models used by the agent. Together, limited capabilities and enriched models, enable the agent to be social. The resulting fractionation matrix can be used for analytic purposes. We use it to examine two social theories—Festinger's Social Comparison Theory and Turner's Social Interaction Theory—to determine how social such theories are and from where they derive their social action.

The social sciences assume that humans are inherently social agents, but "socialness" is open to many views. Consider some characterizations of people: Rousseau's noble savage, *homo economicus*, Skinner's contingently reinforced human, Simon's boundedly rational human, the imperfect statistician, Mead's symbolic human, Blau's human as a bundle of parameters, human as a social position, *homo faber* (human the toolmaker), *homo ludens* (playful human). And no doubt others, especially as more ideology is permitted, e.g., the one-dimensional human of Marcuse (1964).

*This research was supported by the defense Advanced Research Projects Agency (DOD), and monitored by the Avionics Laboratory, Air Force Wright Aeronautical Laboratories, Aeronautical Systems Division (AFSC), Wright-Patterson AFB, OH 45433-6543 under Contract F33615-87-C-1499, ARPA Order No. 4976, Amendment 20.
The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. government.
The authors would like to thank Sara Kiesler and Randy Collins for their comments.

In this paper, we attempt to construct a definition of the social agent—a Model Social Agent.[1] We aim for a conception that is rich enough to do justice to diverse views, but is precise enough to support a rigorous science of social systems. We develop our definition by asking the question, "What is necessary in order to build an adequate artificial social agent?" To some social scientists this approach may seem bound to become trapped in the simplicities of machine-like systems. To us, exactly the contrary seems to be the case. Current theories of cognition, based on information-processing notions, are in many ways (though not all) rich rather than impoverished. Moreover, these notions provide an analytical tool to ask critically *what more is needed to attain social behavior*? These tools allow us to peel away what is understood about individual cognition, narrowly conceived, so as to reveal wherein lies what is social.

We now foreshadow our analysis of the Model Social Agent. We describe the social agent along two dimensions: processing capabilities and differentiated knowledge of the self, task domain, and environment. We fractionate the agent's characteristics along each dimension. This fractionation results in an iterative set of models, each more like the social agent required by the social sciences. Fractionation of the processing dimension reflects limitations to the agent's capabilities that enable it to be social. Fractionation of the knowledge dimension reflects the richness of the agent's perceived environment that evokes and supports social behavior. As we move closer to the social agent required by the social sciences, the agent's information-processing capabilities become more limited and the agent's knowledge becomes more complex (both in type and in quantity). The resulting Model Social Agent is the end point of the two sequences (capability and knowledge). For analytic purposes, the sequencing along the dimensions and the resulting fractionation matrix may be as important as the final social agent. The fractionation matrix indicates what sorts of social behavior arise at various levels of abstraction and how to proceed to create a more adequate model of the social agent.

First we present the fractionation of agent characteristics in a straightforward fashion. We give this sharp analytical form, by grounding it in a theory of the human cognitive agent that is embodied in a specific information-processing system called Soar (Laird, Newell, and Rosenbloom, 1987). Then we turn to some applications of the fractionation matrix thereby illustrating its usefulness and power. We conclude by commenting on how far current models are from an effective Model Social Agent.

We do not claim perfection for this analysis. We focus primarily on the goal seeking nature of agents and assume that they have human sensory and motor capabilities. Important aspects remain on the agenda for further research. Two such issues are personality and motor control. For example, how do motor activities affect social

[1] A similar device has been used with some success in the area of human-computer interaction (Card, Moran and Newell, 1983, Ch. 2), called the *Model Human Processor* (MHP). The Model Human Processor is an attempt to provide a theoretical model of the individual human user that designers of computer interfaces can use to think about human-computer interaction. MHP is based almost entirely on cognitive psychology and is an attempt to integrate into a single framework many experimental results on reactions, uncertainty, memory, learning, and so forth. A Model Human Processor for human-computer interaction is much simpler than a Model Social Agent for the social sciences, which is what this paper addresses. But that does not gainsay the benefits to be gained from attempting the latter.

behavior. However, even in its present form, the analysis appears to have its uses. It provides a framework for the discriminative incorporation of results from artificial intelligence (AI) and cognitive science, by showing that many characteristics of social agents also hold for more general agents. It reveals that many seemingly social theories, even those which rely on situated action, are largely theories of nonsocial behavior—which is not to disparage them, but only to show the source of their power. It permits seeing more clearly the relations among the plethora of views social scientists have espoused regarding the nature of humans. None of this constitutes the ultimate yield, which should be to provide the model of the social agent that enters into theories at the social level. However, this last requires sustained positive theory construction, whereas all we can provide here is introduction, analysis, and perspective.

## THE MODEL SOCIAL AGENT (MSA)

The Model Social Agent has information-processing capabilities and knowledge. Agents' information-processing capabilities are goal oriented. They control the agent's ability to handle information. Agents exist within an environment which is external to the agent's processing capabilities. The agent's knowledge is to an extent dictated by the external environment in which it is situated. The Model Social Agent exists in a particular situation (both physical and social). This situation is the environment perceived by the agent, but how the agent encodes it, and how much of the environment is encoded by the agent, is an open issue. The agent has a goal. The agent enters a situation with prior knowledge. The agent may have an internal mental model of the situation that differs from that held by other agents int he same situation. Throughout, we take the agent as having the typical human sensory and motor devices to sense the environment and affect the situation. We are concerned with the nature of the inner system that makes interaction social, not with how the sensory and motor capabilities affect social behavior. Whether specific sensory and motor capabilities are requisite for social behavior is an interesting question, but it is not ours.

The distinction between information-processing capabilities and knowledge can be illustrated by considering artificial intelligence (AI) systems. Consider a system for playing chess. Encapsulated within the capabilities may be the rules for playing chess, "motor" procedures for moving chess pieces, procedures for encoding the board and analyzing the agent's position, and goals (e.g. to win). However, simply having the capabilities does not enable the agent actually to play chess. In order actually to play chess the agent needs knowledge. In this case, the requisite knowledge includes information about a specific partner, board layout, and the moves made by the partner. The agent's behavior is determined jointly by information-processing capabilities and knowledge, differences in either will lead to differences in behavior.

Subdividing the two dimensions, information-processing capability and knowledge, results in the *fractionation matrix* displayed graphically in Figure 1. Taken together, the two fractionated dimensions result in a matrix of possible agents in

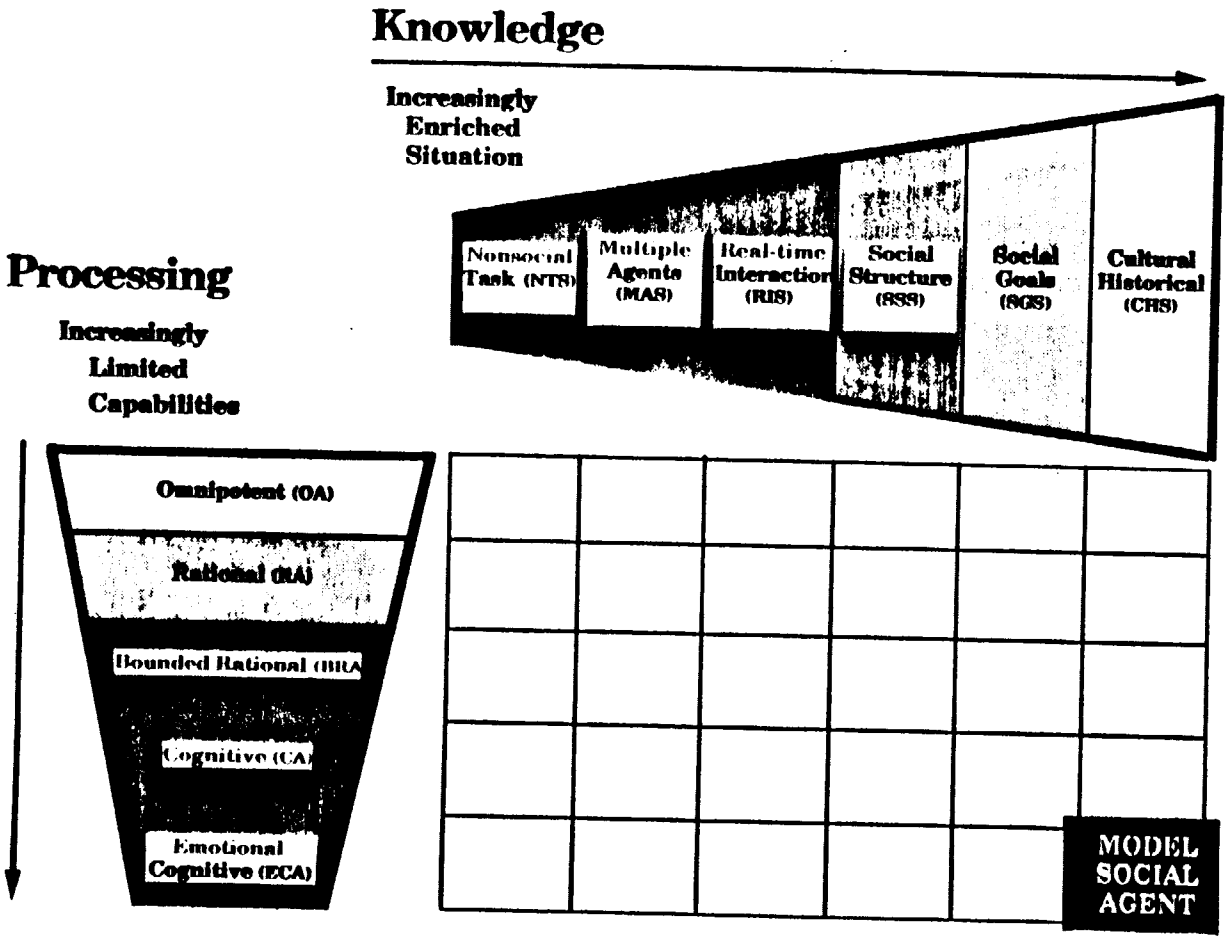


FIGURE 1. Model social agent.

which we can locate the Model Social Agent. As shown in Figure 1, the Model Social Agent is minimally capable and maximally knowledgeable. By maximally knowledgeable we mean that the agent is in an enriched situation and has considerable knowledge corresponding to the situation.

The process sequence starts with a maximally capable agent and successively restricts its information-processing capabilities, obtaining a more and more realistic theory of a human being. As information-processing capabilities are reduced, the agent becomes less capable of achieving its goals. This results in a hierarchy of agents (as in Figure 1) in terms of performance; i.e., the omnipotent agent can do anything the rational agent can do, the rational agent can do anything the boundedly rational agent can do, and the boundedly rational agent can do anything the cognitive agent can do, and so on. Each successive agent refines the previous agent with yet further limitations. The refinement reflects additional details on the information-processing mechanisms required for the previous agent. As the capabilities of the agent decrease, task performance may degrade, but different, and possibly more complex, behaviors emerge. This is because a more capable agent may not have need of certain behaviors, e.g., the omnipotent agent does not need to gather information. It is important to keep these two aspects distinct—more limited performance but richer behavior.

Emotions enter into this sequence because emotions affect processing capabilities.[2] Emotions cannot add information processing capabilities, but they can degrade or limit current capabilities. This limitation can affect goal attainment. Thus, for each model of a limited agent, another agent can be envisioned that is limited further by having emotions. The basic affect of the emotions is the same on each of the models, differing only because of the differing abstraction of each processing model. Hence, we will introduce emotions only once, in connection with the model of the cognitive agent. At that point, we will comment on how emotions operate as applied to the less limited agents.

All agents, regardless of their information-processing capabilities, have goals, the ability to change goals, and the ability to interact with other agents and objects. However, the situation and the agent's knowledge of the situation specifies what goals, whether the agent actually changes goals, and what other agents or objects are present to interact with. Moving from right to left along the knowledge dimension in Figure 1 the environment characterized becomes a more abstract version of the external environment in which the agent actually is situated. The knowledge sequence, moving from left to right, starts with a minimally knowledgeable agent and successively enriches the agent's knowledge. The agent's internal mental models become a more and more realistic theory of the human social environment. The stages of this sequence are expressed as situations. However, this defines a corresponding sequence of agents, namely, those with mental models (the internal knowledge and organization than enable operation) of the stipulated environments (Figure 1). Increasing the richness of the situation corresponds to the agent increasing the complexity and detail of the information available to the agent. The agent

[2]Emotions also may reveal to others the goals of the agent, or may even affect which goal the agent currently is working toward. These functions are distinct from the role of emotions in limiting information processing.

who knows this information has less abstract mental models. Knowledge is cumulative across environments in the sense that an agent in a less abstract environment has knowledge from the more abstract situation enriched with further information. As the agent's mental model is enriched the agent acquires a wider repertoire of actions and a more complex set of goals. Agents with more abstracted mental models have fewer actions and fewer, or at least, less complex goals. Without certain knowledge certain actions are neither necessary nor possible. For instance without knowledge of the social structure the agent does not need actions for locating its position within the social structure nor can the agent locate its position. There is an interaction between capabilities and knowledge. We are not claiming of the agent (other than the omnipotent agent), that the agent in a particular situation knows all there is to know about the situation. In terms of knowledge, the omnipotent agent is, given a particular situation, omniscient. For all other agents, how much of the situation the agent knows, and how it learns it, are governed by its capabilities and any physical or social constraints imposed by the situation. Rather, what we are claiming is that as we take the same agent, e.g., the cognitive agent, and move it from left to right, the types of knowledge that are considered by the agent increases.

It is important to recognize that there is a social and physical reality external to the agent. The physical reality requires an explanation in its own terms. The social reality emerges from the on-going interactions among social agents, exists as a social entity that can in turn constrain future interactions, and can be altered by these interactions. In this paper, we do not seek to provide an explanation for either the physical or social reality, nor do we define the processes by which the social reality is constructed and constrains human activity. Rather, we argue that in order to truly understand the relationship between social and physical reality and human agency, one must have an adequate Model Social Agent in order to understand the emergence and alteration of social reality. Whereas, the existence of social reality as a social entity is external to the agent model in the same way that the physical reality is external to the agent, and thus may be governed by its own laws. Whether or not there are such laws is beyond the scope of this paper. Of import here is that as we move along the knowledge dimension the agent is gaining increasing knowledge of these realities, and so ability to respond to them.

A final caveat on the knowledge dimension. As we move along this dimension the type of knowledge changes from knowledge about the task, to knowledge about the current society, to knowledge about the society's culture and history. No claim is being made that the agent knows everything associated with each situation. Further, no claim is being made that the agent's knowledge is "complete" rather than generated. Rather, the claim is simply that an agent at a particular level of abstraction has knowledge of a certain type and that all more complex information has been abstracted out of the agent's mental models.

Capability and knowledge are distinct dimensions and we can consider refinements in each dimension separately. Moving along either dimension generates a theoretically richer model of the agent, a closer approximation to a model of a social human being. The agent simultaneously at the end of both dimensions is a useful candidate approximation to a social human being (Figure 1). We call such an agent the Model Social Agent. As we move out on these dimensions, the issue is not to

judge each agent against the human in some absolute way. Rather, we assert that models further along on either or both dimensions should be treated as a successively better approximation to the human, capable of exhibiting certain behaviors of a human, but not others. It seems to us that everything is needed for a Model Social Agent capable of exhibiting all social behaviors. Nevertheless, important social behaviors may emerge at earlier stages. Of course, this is what is to be discovered ultimately.

### Increasingly Limited Processing Capabilities

We now describe a set of agents by moving along the processing dimension. As we iterate through the information-processing capabilities we will be altering the agent's capabilities but not its knowledge. Thus, each agent we describe along this dimension can be thought of as knowledge independent. We abstract away all knowledge of the environment. We do this in order to illustrate that successive limitations on the agent's information-processing capabilities, independent of what knowledge the agent has, limit the agent's ability to attain its goal. Processing limits do not preclude nor make possible specific behaviors on the part of the agent; however, as will be seen, processing limits may make certain behaviors necessary. Each of the agents we describe should be viewed as a class of agents. Agents within this class would vary in knowledge; i.e., how abstract their mental models of the world are, but not in capabilities.

#### The Omnipotent Agent (OA)

The starting point is an agent who is completely capable; i.e., who is *omnipotent*. The omnipotent agent will, given the situation in which it is embedded, be omniscient with respect to the knowledge germane to that situation. This is an agent who knows all there is to know about the task environment in which it is embedded. The agent's behavior is determined by the nature of the actual task environment plus the goals of the agent. For example, if the omnipotent agent is within a nonsocial task situation that agent knows all task related knowledge but has no knowledge of other agents or the underlying culture. The agent will take whatever actions are possible and necessary to attain its goals. The agent can take actions to change its environment. The omnipotent agent is, of course, the original *homo economicus*. Within economics this agent has been extended to task environments that are uncertain and so became the initial rational expectations model. Characterizing the environment as objectively uncertain does not change the omnipotent status of the agent, who still knows all there is to know given a particular situation. The success of economics shows just how good an initial approximation this agent is, especially when used to analyze certain types of large institutions such as markets.

#### The Rational Agent (RA)

The first limitation comes from recognizing that an agent does not know everything about its task environment. An agent has its own body of knowledge and it behaves

rationally[3] with respect to that knowledge by taking those actions that its knowledge indicates will lead to attaining its goals. The task environment is largely mediated by this knowledge, but still influences the actual actions that can occur. We leave *task environment* in Figure 1, to indicate that the agent's knowledge is about the task environment, hence an analyst's knowledge of the task environment offers an approximation to what the agent knows.

This model of an agent is called a *knowledge-level system* in computer science (Newell, 1982).[4] It is an abstraction of an information-processing system that has sufficient time and adequate methods to exploit all of the knowledge it has acquired about its environment. The knowledge-level system abstracts from the way the knowledge is represented and from the processing that is required to extract from this representation the knowledge about how to act. All that remains is the content. Besides restricting the performance expected of an omnipotent agent, the rational agent adds an essential activity missing from the latter, namely, the acquisition of knowledge. No such behavior makes sense for the omnipotent agent, which in effect already knows all there is to know within a given situation. The rational agent, which in effect already knows all there is to know within a given situation. The rational agent, like the omnipotent agent has actions for altering the environment. In addition, the rational agent has perceptual or knowledge-input actions for interacting with the environment. From a more sociological perspective, rational agents exhibit the locality so dear to organizational theorists. That is, the agent's knowledge is dependent on the places it occupies in the organization (or society). Despite having limited capabilities, the knowledge-level system or rational agent is still very much an idealization.

### The Boundary Rational Agent (BRA)

The concept of bounded rationality was introduced by Simon (1957, 1979, 1983, see also March and Simon, 1958) and has become familiar throughout the social sciences. The boundedly rational agent has limited attention and therefore cannot process all the knowledge available in its task environment. As Simon (1976) noted, this agent is *procedurally rational* in that it reliably deploys its processing capabilities to attain its goals. But its attempts to do this are limited by its abilities and knowledge.

The stipulation of a computationally-limited agent does not specify the form or nature of the limits. Early on, the concepts drawn from computation were general

---

[3]Substantial confusion in the social sciences is caused by the term "rational". At times, rational is used to refer to "procedural rationality" (Simon, 1976); i.e., an agent is procedurally rational if given the same information at two different points it will produce the same solution. At times rational is used to refer to task rationality; i.e., using only the information directly related to a task and not using external information such as social and cultural information. Typically, the rational agent is characterized as being both task and procedurally rational. In contrast to these views, currently in cognitive science the term rational often is used to mean "bringing all the information the agent has to bear on the problem, regardless of the source of the information or the type of information." For further discussion, see Wuthnow (1988: pps. 489–490).

[4]Following economics, the omnipotent agent might be taken to define the rational human. Then this model of actor might be called the knowledge-level agent. But rationality does not have to do with *what* is available about the environment, but with the response to *whatever* is available. So this seems a better choice to identify as the rational agent.

TABLE 1
Tenets of the Human Boundedly Rational Agent

1. Humans decompose tasks into goals and subgoals.
2. Humans encode task environments into internal representations.
3. Humans conduct searches to find information or solve problems.
4. Humans can respond cognitively within a second.
5. Short-term memory is of limited size (7 + / – 2 chunks).
6. Long-term memory is associative and hierarchically organized by chunks.
7. Chunks are built every couple of seconds (long-term memory acquisition) in an automatic rather than deliberate fashion.

and qualitative, such as *programmed* vs. *unprogrammed* behavior (March and Simon, 1958). But with continued developments in computer science, artificial intelligence, and cognitive science, a more exact picture has emerged of the general nature of the limitations (Table 1). The boundedly rational agent has the following features: The agent is engaged in one or more tasks. To perform this task the agent creates an internal (or cognitive) representation of it. The cognitive structures required are a *long-term memory* for permanent knowledge and a *short-term memory* for the immediate situation. There is a unit of memory organization, the *chunk*. What chunks the individual has are developed over time. These chunks are not all accessible immediately but are linked together in some associative hierarchical structure that must be searched in order to solve problems. This search is directed by the goals and subgoals of the task faced by the agent. Finally, thought, which includes problem solving and learning, takes time.

The human, of course, is much more complicated than this. However, the characterization of the agent provided in Table 1 is sufficient for reasoning about boundedly rational behavior in many situations. Indeed a wide range of studies using agents with some, although rarely all, of these characteristics have enhanced our understanding of human behavior, particularly in the realm of organizations.

## The Cognitive Agent (CA)

The cognitive agent is a refinement of the boundedly rational agent based on an understanding of human rationality that goes beyond Table 1. The boundedly rational agent is characterized by the set of tenets or principles described in Table 1. These principles characterize important properties of human rationality. They permit many inferences about social behavior that are beyond more abstract rational and omnipotent agents. But these principles are incomplete. They do not describe perception, motor control, the impact of interruptions on behavior, the separation of intention from action—or any of a number of other aspects of human behavior that can have consequences for social action.

To describe all aspects of human behavior we might increase the number of principles. However, so many would be required that we might be better off if we specify

an *architecture*.[5] An architecture is a fixed set of information-processing mechanisms that are used to generate all behavior. These mechanisms embody all processing limits. From the standpoint of creating an artificial social agent, an architecture is necessary in any event. Table 1 is too sketchy to act as a specification for an actual agent.

The cognitive agent is the boundedly rational agent with a fully specified architecture. We can think of the cognitive agent as the culmination point of all information-processing limitations on the boundedly rational agent. The cognitive agent does not introduce qualitatively new categories of limitation as did the higher agents—adding the category of knowledge limits to obtain the rational agent, and the category of representation and processing limits to obtain the boundedly rational agent. Where the boundedly rational agent is a set of general claims the cognitive agent is a set of specific operational details. Moving from general principles to specific architecture further limits the agent. Where the boundedly rational agent may have some type of knowledge structure called a chunk, in the cognitive agent we know the exact form of these chunks and the procedure for creating them.[6]

We take as our cognitive agent—Soar (Laird, Newell, and Rosenbloom, 1987). The Soar architecture stems from an ongoing effort to integrate cognitive science findings into a unified theory of cognition (Newell, 1990).[7] Soar, as cognitive agent, operates by continuous recognition in an internal field of salient knowledge, making microdecisions several times a second to shape the flow of its considerations. This internal cognitive world operates in a loosely coupled fashion with the bodily systems of perception and action.[8] Cognition must operate with the mixture of knowledge provided by perception and elicited experience. But it can engage in planning and imaginings that are decoupled. Similarly, perception and action proceed at their own timescales. Cognition arrives at intentions which then have to modulate the ongoing activity. We continue by describing only those aspects of Soar necessary to understand the differences between the boundedly rational agent and the cognitive agent.

Table 2 lists the key properties of the cognitive agent as embodied in Soar. Figure 2 shows how the components of the cognitive agent fit together. To summarize: (1) the cognitive agent's working memory changes as the agent makes decisions and gathers information through its perception; (2) cognition acts in a supervisory fashion to define what actions the agent takes; and (3) the learning mechanism (the chunker) creates new rules for the agent which are then stored as part of long

---

[5]Within computer science the term *architecture* refers to the fixed structure of a computer.

[6]Of course, by being more specific, the cognitive agent described here has more risks of being wrong. The generalization about human behavior in the boundedly rational agent are by now quite secure empirically.

[7]Soar plays other roles as well. It is being used to form a second iteration of the Model Human Processor, the effort mentioned earlier for the field of human-computer interaction. Also, it is being used for artificial intelligence research into learning systems and expert systems. These roles are mutually compatible. We note them here because we will be treating Soar only from the perspective of the fractionation scheme.

[8]Although the exact nature of the relation between cognition and the perceptual-motor system is not very clear yet, we find it useful to think of the cognitive part of the agent as having supervisory control over the perceptual-motor system.

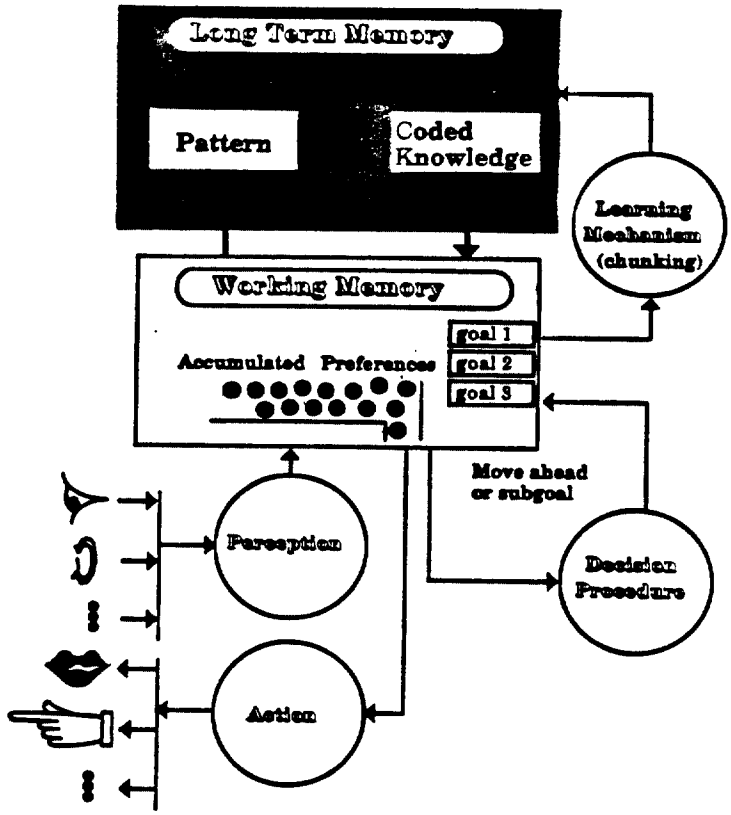


FIGURE 2. The structure of the cognitive agent.

term memory. Many of the details of Soar are not described,[9] but collectively they demonstrate that all the properties set out in Table 2 are embodied in a single operational system that is generally intelligent. Soar has many of the characteristics of human cognition, such as performance speed ups with practice on skilled tasks, errors (like human errors) in speed reasoning tasks, varying response time with the complexity of the mapping of inputs to outputs, and strategy changes in problem solving (Lewis, Newell and Polk, 1989; Newell, 1990; Polk and Newell, 1988; Polk, Newell and Lewis, 1989; Ruiz and Newell, 1989). These properties should emerge from the architecture of the cognitive agent.

Soar, as cognitive agent, holds acquired experience in symbolic coded form and enables the encoded knowledge to be used in doing tasks. Soar's architecture (see Figure dca) has two memories and four mechanisms: *long-term memory* for permanent knowledge, a *working memory* for temporary knowledge about the current situation, a *decision process* that uses immediately available knowledge, a *learning mechanism* (i.e., the chunker) that creates permanent knowledge from momentary experience, a *perception system* that encodes the external world, and an *action system*

[9]For further operational details on Soar refer to Laird, Newell and Rosenbloom (1987), and Laird, Rosenbloom and Newell (1986a, 1986b). For a more generalized view of Soar see Waldrop (1989a, 1989b).

TABLE 2
Key Properties of the Cognitive Agent (CA)

1. Goals
   - Goals direct action
   - Goal stack can change over time

2. Components of the architecture
   - Long-term memory
   - Working memory
   - Decision procedure
   - Learning mechanism (chunking)
   - Perception
   - Action (external)

3. Performance
   - Provided by a sequence of microdecisions (several per second)
   - Salient knowledge continuously and automatically becomes available
   - Candidate decisions and preferences about these decisions accumulate
   - Agent decides according to whatever preferences are available
   - If the agent cannot decide, it automatically sets up a subgoal to obtain knowledge to be able to decide

4. Awareness
   - Engages in both automatic and deliberate actions
   - Long term memory is a recognition memory

5. Learning
   - Experiential learning occurs every few seconds
   - The agent automatically records the knowledge that resolved an impasse with the pattern to re-evoke it (a chunk)
   - Chunks become a permanent addition to long-term memory

6. Interaction with the world
   - Behavior is organized by problem spaces
   - Perception and action systems run independently of cognitive system
   - Perception continuously adds coded knowledge to the working memory
   - Cognition provides the intention for action by deciding on what commands to make to the action system
   - Interrupts occur as the situation changes and the agent must respond to the incoming information rather than continue to only utilize information in long-term memory.

that decodes internal commands into interactions with the external world (includes speech and writing). The mechanisms, and also the working memory, are fixed devices. The long-term memory, on the other hand, expands indefinitely to accommodate all the knowledge that accumulates from experience. The long-term memory is active and associative. The agent uses the knowledge it has accumulated through experience to perform tasks such as evaluating its position in the group, determining whether to take place in a social movement, interacting with others, playing games, or reading stories. This knowledge includes the agent's experience of itself as a physical, biological, and social organism.[10] Through the perception and action

[10]Nothing in this is meant to choose sides on various nativist-empiricist controversies. The cognitive mechanisms are purely biological, but the knowledge encoded in the system may come from either biological or experiential sources.

system the cognitive agent interacts with the physical and social reality. Whether in fact the social reality emerges from the ongoing interactions among a set of agents (cognitive or MSA) or exists independently governed by its own internal logic is an issue separable from how the agent interacts with that social reality. By specifying how this interaction occurs, as is done for the cognitive agent (through the Soar architecture), it is possible to begin to address whether, given a collection of cognitive agents social reality is a completely emergent phenomena. We would argue that the controversy around individualistic basis for social behavior can be addressed much more fruitfully by examining the behavior of collections of fully specified cognitive (or better yet, MSA) agents.

The cognitive agent's behavior is governed by the fact that the agent is continuously making a series of highly diverse microdecisions. These microdecisions affect the internal course of thought and occur a few times a second. This may seem fast, but think of how fast a person detects the point of a joke, or judges someone to be lying by a slight turn of phrase and hesitation. This microdecision procedure enables the agent to see relations, draw inferences, make considered judgments, and determine new preferences. These larger cognitive operations result from multiple microdecisions; the microdecision itself simply works on the current cognitive state to determine what to think in the next instant (quarter second). Once a specific microdecision is made, the agent simply takes it and moves on to the next. Cognitive life is thus a continuous flow of microdecisions supported by a continuous flow of knowledge from long-term memory. Importantly, this all happens automatically in the cognitive agent. The microdecisions constitute the process of awareness—they are the finest-grained points at which the agent can be deliberate. The processes that make them up—the inflow from long-term memory and the decision process itself—are beneath awareness. Likewise, all the available coded knowledge that has poured into working memory but is not being attended to by the decision process lies outside awareness. The larger deliberate goals for the agent arise from coded knowledge in long-term memory. But the moment by moment flow of cognition is governed by the automatic subgoals continuously arising and getting resolved in the attempt to make microdecisions—which of course are themselves occurring in the attempt to decide on external actions or arrive at judgments based on extended considerations, and so forth. Because of its automatic nature, the agent cannot necessarily articulate these subgoals. Indeed, articulation is an external action which occurs only if the knowledge can be brought to bear to find the right things in working memory and issue the appropriate action commands. This basis for awareness, and inability to articulate goals, move the cognitive agent beyond the boundedly rational agent.

In Soar (and hence the cognitive agent), the flow of cognition is in the service of interaction with the external world. There is a flow of external (coded perceptual) and internal (long-term memory) knowledge into working memory to make up the total available salient knowledge. While we have assumed that all agents have sensory and motor capabilities, it is only when an architecture is specified (as it is in the case of cognitive agents, such as Soar) that details on the relationship between the various components and cognition come into play. Attending to these relationships in the detail necessary to specify an architecture brings to light constraints on

human behavior beyond the more generic constraints of bounded rationality. For example, the cognitive agent (as Soar), can be described as attending to the external situation, with long-term memory automatically elaborating, embellishing and interpreting the scene. On the action side, codes for commanding external action flow from long-term memory into working memory as actions-to-be-considered. Microdecisions provide the *intention* to take a specific external action. This releases these command-like codes to affect the motor system, which then moves in a new way. Motor actions take at least seconds, which is longer than microdecisions. Thus, there exists for the cognitive agent (unlike the boundedly rational agent) a separation between intention and behavior, which opens the door for all the conundrums of responsibility and attribution of personal causality. The above structure permits the separation of thought and action—that the agent can actually say "hello" or only think hello.

We have been treating long-term memory in the cognitive agent as if it were a fixed repository. In fact, in Soar, as cognitive agent, new knowledge is being added continuously to long-term memory in the form of new productions or "chunks". All new permanent knowledge arises from experience. This learning mechanism (called *chunking*) is a way of continuously formulating past experience for storage so that it can be re-evoked when appropriate. It is not based on an inductive assumption or predictions about the future. If tomorrow is like today, it may get evoked and be useful; if tomorrow is different, it simply will not get evoked.[11] For the cognitive agent, unlike the boundedly rational agent, all learning occurs automatically and frequently (every second or so). The agent does not decide to store away various bits of knowledge, and it does not know explicitly what has been stored away.

We have emphasized the temporally microscopic character of the cognitive agent. This is, in fact, the scale at which the fixed architectural mechanisms operate. The scale at which social behavior occurs seems to be much longer—hours, days, weeks, and so on. But the flexibility of the cognitive agent arises because its time grain is fine enough for extended considerations to take into account situational knowledge mixed with its own accumulated knowledge. Consonant with this is that the socially significant act, e.g., the nod of acceptance or the curt turning away, can occur almost instantly on the basis of immediately presented knowledge.

In addition, we have emphasized that the agent proceeds in many respects automatically and without awareness. This is because the fixed architectural mechanisms are exactly the locus of such automatic behavior. Awareness is not a primitive concept, but arises from the ability of agent to base its actions on what it knows about itself and its own operation. This occurs at the level at which microdecisions that reflect such considerations can be taken. This is above the microdecision level that is comprised of the (primitive) processes of knowledge accumulation and preferencing. Awareness occurs only at the macro and not the microdecision level. As

[11]This description is highly abbreviated and may be far away from the reader's experience of computational mechanisms. It is thus important to re-emphasize that Soar has such a learning mechanism, that it operates essentially as described, and that it demonstrates the feasibility of such a scheme. Beyond Soar itself, the field of machine learning provides a substantial body of theory and experimentation about such learning mechanisms that makes clear why this scheme works (see the journal *Machine Learning*).

knowledge is learned and chunks created, the agent may become unaware of the rational behind certain behaviors.

Soar, as cognitive agent, is more limited than the rational agent in its inability to get complete access to all the knowledge encoded in its long-term memory. Soar's processes proceed in some ways automatically and not in accord with what a god's-eye view reveals could really solve its problems. The automatic nature of Soar, as cognitive agent, moves us beyond the boundedly rational agent, where all actions are more deliberate. Within these limits, Soar, as cognitive agent, is capable of rational considerations in the pursuit of its own ends, of responding rapidly to changing circumstances and knowledge, and of learning from its experiences. Soar is a generally adaptive system. While the boundedly rational agent, may or may not be adaptive, the cognitive agent, as embodied by Soar, is limited by the use of specific learning mechanisms (in this case chunking).

### The Emotional Agent

Even a cursory examination of social science literature, not to speak of the world's literatures and the social world itself, establishes that emotions, affect and feelings[12] are an essential dimension of human social behavior. Unfortunately, while there has always been a substantial stream of work on emotions (for general overviews see Izard, 1972, 1977 and Kemper, 1987), there has been little mutual contact between theories of rational and cognitive humans and emotion-based theories of humans (Norman, 1981). A major conceptual gap in cognitive science is the lack of a notion of emotion that integrates with the collection of cognitive mechanisms, such as those that comprise our cognitive agent.

The emotional agent, while more human than the cognitive agent, also is more constrained in its ability to attain its goals. We could add emotions to each of the other limited agents, the rational, boundedly rational and cognitive agents. These all would reflect the same notions of how emotions limit goal attainment. Here, we will focus on the emotional-cognitive agent, as the one with the most detailed considerations, and then come back to emotional-rational and emotional-boundedly-rational agents at the end of the subsection.

Fractionating the total social agent lets us ask: *How does emotion modify and limit the behavior of the cognitive agent*? This shift in perspective is not minor in its effect. Recent attempts to characterize emotions (Frijda, 1987) find themselves assimilating to the emotional system immense amounts of cognitive apparatus. If emotion sui generis is taken as the starting point, an overwhelming need arises to bring in appraisal, action tendencies, behavior control, and so on, all categories that are central to cognition per se.

At this point we cannot put forth a detailed theory of emotion in cognitive agents. However, we can list the major phenomena of emotion that *change* the picture of the agent obtained just from information-processing considerations. These phenomena are listed in Table 3. We start (E1) by taking an emotion to determine an *orien-*

[12]The term *emotion* will be used to cover the phenomena of emotions, affect and feelings. No single term does the job adequately and no theory-neutral distinctions support separate subdomains for emotions, affects and feelings.

TABLE 3
The Phenomena of Emotion That Go Beyond Cognition

| | |
|---|---|
| E1. | **ORIENTATION.** An emotion is oriented to a target determined by cognition. |
| E2. | **POSSESSION.** An emotion takes hold of its possessor. |
| E3. | **COGNITIVE INFLUENCE.** Any component of cognition that can vary in consonance with an emotion will be affected: |
| | E3.1. **ATTENTION.** What aspects of the world are attended to. |
| | E3.2. **APPRAISAL.** How these aspects are evaluated. |
| | E3.3. **ACTION.** What actions are selected and how they are performed. |
| E4. | **INTENSITY.** The greater the intensity the more consonant are the effects. |
| E5. | **DIVERSITY.** Emotions may have any targets determinable by cognition. |
| E6. | **PERSISTENCE.** Targets continue to evoke their emotions. |
| E7. | **HABITUATION.** Repetition of the occurrence of an emotion decreases its intensity. |

*tation* toward some events, agents or objects—angry at $X$, gratified over $Y$, fearful of $Z$, warm and cozy with respect to $A$, and so forth. *Events, agents and objects* is the taxonomy used by Ortony, Clore and Collins (1988) in their recent treatment of the cognitive structure of emotions; for short, we refer to any of these as the *target* of the emotion. We hedge (necessarily, given the discussion above) about what it actually is in the agent that produces this orientation. Having an orientation towards a target implies that something is identified and delimited to be the target. Such capabilities are exactly what cognition provides and are not posited anew for emotions. However, making that move—emotions are with respect to targets *as determined by cognition*—already constrains what emotion (or an emotion system) can be. We hedge again in the term *orientation*. Operationally, an action, intent or appraisal concerning a target is *consonant* or not with respect to an orientation, and to varying degrees. For example, hitting is consonant with anger, hugging not (ceteris paribus). The point here is not whether observers can be accurate in their judgments about consonance in actual situations, where all the complexity of intentions and hidden goals enter in (the hug is used to break the ribs); the point is that an orientation provides the framework for attempting such judgments.

The key feature is E2: emotions take hold of their agents independent of the mental state, given that the targets are identified. Emotions are outside of cognition, which neither bids them come nor banishes them at will, so to speak. This is what makes the emotional agent different from the pure non-emotional agent—there are independent sources of influence at work in the inner chamber of information processing. What actually differs behaviorally, of course, depends on how emotions interact with cognition. The frame of reference for that interaction is the array of targets that cognition presents to emotion.[13] Given an array of targets, emotions arise. Each determines a notion of consonance with the emotion; i.e., it determines an orientation against which aspects of cognition are more or less consonant. Aspects of cognition that permit such a relation are subject to be influenced (E3). Chief among these are: attention (the aspects of the environment on which the agent is spending its time, either on or off the target, depending on the emo-

[13]Thus, emotion seems completely dependent on cognition. But this is only half the story. Since emotion affects cognition, it influences the landscape of targets that cognition presents to emotion. Thus the influences are mutual and potentially intricate.

tion); appraisal (does the appraised entity increase or decrease the consonance) and action (does the action increase or decrease the consonance). The direction of influence always is toward more consonance, and the intensity of an emotion is the extent of that influence (E4). Further, given an array of targets (presented by cognition) the effect of the emotion may be different (love manifests itself differently for different objects) (E5). Finally, emotions continue to be evoked as long as their targets exist (E6), although there is habituation; i.e., the more often an emotion is manifest the lower its intensity (E7).

Some essentials are missing from this picture. As described, it appears to be like a continuous field, in which the actual information processing is the result of all the emotions that tug on it in their own direction of consonance and with their own intensity. We have not shown how such a continuous field actually works; presumably this is intimately related to what sort of things emotions are within the mental system. Also, we have not described the exact nature of an orientation. Is it a symbolic structure of some kind, separate from or intermingled with the symbol structures that represent the target? Is it a set of operators for processing targets to yield consonant symbol structures? The picture certainly does not settle many issues about emotions (e.g., whether there exists a primitive set of emotions). However, it does provide an explicit enough statement of how emotion relates to cognition to justify a separate fractionation for an emotional-cognitive agent. In addition, it has some strong implications, e.g., emotions affect cognition, not directly behavior, and cognition affects emotion in part by determining the target array. Besides the highly dynamic effects of mutual influence, this latter opens the door to the elaboration of emotions by learning.

Even this minimal discussion of emotions illustrates that emotions serve to further limit the agent's capabilities. In a limited-knowledge agent, emotions affect the agent's ability to attain its goals by further affecting the agent's locus of attention, ability to evaluate information, what actions are taken, and so what is learned. This may have positive or negative effects on goal attainment. The omnipotent agent has risen above having even limited knowledge. Emotions, in this fashion, cannot dampen its ability to attain its goals. Emotional considerations can be added to the cognitive agent, the rational agent, and the boundedly rational agent. The ability of emotions to shape and restrict attention, appraisal and action can be formulated in terms of shaping and filtering the knowledge that a rational agent has, while still preserving that the agent's knowledge determines its actions. This fails to include many aspects of emotion of course—the emotional-rational agent remains somewhat bloodless—but that is the price of the knowledge level of abstraction. In addition, for the emotional boundedly-rational agent emotions alter the forms of representation used, restrict working memory size, and so alter what is learned. While we do not know exactly how emotions will restrict the cognitive agent we expect that emotions might further limit the cognitive agent in the following ways. Emotions should take hold of the agnet (Table 3: E2) and reduce performance by slowing down the decision cycle and rate of learning (Table 2: 3 or 5). Further, emotions, by affecting the agent's attention and appraisal capabilities (Table 3: E3), affect the agent's ability to interact with the world (Table 2: 6) by making some objects or events more salient than others. Details on the mechanisms by which emotions and cognition

inter-relate await further study. We do expect, however, that some of these mechanisms will center on the way in which emotions can set deliberate goals for the agent.

The emotional-cognitive agent provides an appropriate terminal point for the sequence of agents formed by constricting its processing powers, starting from the super-rationality embodied in the omnipotent agent. However, there might be further limitations, analogous to emotion; i.e., aspects of the human that lie outside cognition and effect processing in ways other than just as additional input. The human is a biological organism and as such it is subject to fatigue, disease, and other biological effects. These are not emotions, but they do (or may, the evidence is not always clear) affect the very nature of how the agent cogitates. Another example is the panoply of phenomena gathered under the rubric of personality. These are surrounded by unclarity, but are certainly important to what constitutes a social agent. There is little doubt about the robustness of some of the phenomena per se, such as introversion-extroversion (Eysenck and Eysenck, 1985) or the propensity for risk taking (Kogan and Wallach, 1964; Tversky and Kahneman, 1979). Almost all research on personality is variable oriented. The same barrier that was faced with emotion exists, namely, relating these variables to cognitive mechanisms. In addition, there is the issue of how much of personality can be adequately handled by the body of knowledge the held by the cognitive agent (i.e., at the rational-agent level) and how much can be adequately handled by the increased capability restrictions arising from the emotional agent. Until these aspects are clarified, it is hard to give proper form to additional personality fractionations and to decide how they fit into the total scheme. So, with the proviso that additional fractions may be possible, we leave the processing-limits dimension.

### Increasingly Rich Situations

As previously noted, the set of behaviors an agent can engage in is defined by the agent's capabilities and its knowledge. We now describe a set of agents by moving along the knowledge dimension. As we iterate through the situations we will be altering the content of the agent's knowledge but not its capabilities. Along this dimension we will proceed from right to left, from the most complex real situation to the most abstract. Rather than describe all the information that gets added as the situation becomes more realistic we describe the type of information that gets abstracted away from as the reality of the situation decreases. Each agent we describe along this dimension is less knowledge intensive than the last. By decreasing the types of knowledge available to the agent we decrease the types of behaviors in which the agent can engage. In describing the agent's knowledge we ignore the agent's capabilities. Regardless of its capabilities, if the agent is too impoverished in its situational knowledge it will not behave as a social agent. Each of the agents we describe should be viewed as a class of agents. Agents within this class would vary in information-processing capabilities, but not in available knowledge.

The issue in defining the knowledge dimension is not how much information the agent must acquire. Whether the agent comes to the situation with only some highly generic knowledge about situations in its long-term memory, as does Soar, or with previously encoded knowledge about the specific situation, as does an expert system,

the situation may still be nonsocial. Nor is the issue one of method. The agent in a nonsocial situation, e.g., the nonsocial task situation, may examine the task to see what methods might work. Rather, the issue is what type of information the agent has, and how that type of information informs action.

The situation plays a somewhat different role with different agents in our processing sequence. For the omnipotent and rational agents, most investigations have focused on understanding how the behavior of the agent changes as the environment or situation is varied. These are differences due to knowledge. The rational agent, as opposed to the omnipotent agent, permits an additional degree of realism into such analyses, by limiting the knowledge the agent has of the environment. But the knowledge and goals ascribed to the agent are still dictated by the external situation. For instance, the goals of production, accounting and sales personnel can be distinguished largely on the basis of their acquired expertise and knowledge of their own organizations. For the more refined agents—boundedly rational, cognitive, and emotional—the focus shifts to a more internal view. There is a correspondence between the external situation and the internal capabilities of the agent. An agent cannot deal with a situation without having acquired the knowledge to do so.

### Cultural-Historical Situations (CHS)

We take as our epitome of the real world the cultural-historical situation. The agent exists at a particular point in time and history, in a particular society, at a particular place, with particular institutions, and so on. The situation is affected by, and has resulted from, a historical process that led to the current state and moment. The agent has available a body of social practice that has been shaped by a specific culture in the ways that may be idiosyncratic to that culture. The agent has "developed" within this situation and so is socialized to a particular way of doing things, to specific beliefs, norms, and values. For example, the socialized agent faced with an event knows the appropriate response, what sanction to expect should the response not occur, and who will suffer should the sanction occur or not occur. Such an agent might eat with a fork but be confused when faced with a pair of chopsticks; might reach out to shake hands when introduced only to be embarrassed if the other agent simply nods; might suffer a sense of moral outrage when pollutants are dumped in the local river but refuse to personally recycle goods.

### Social Goals Situation (SGS)

To reach this situation we abstract away cultural and historical information as an evolving environment. The static artifacts of a live culture remain. The social agent has multiple goals of specific types: (1) task-related goals, (2) self-maintenance and enhancement goals, and (3) social-maintenance goals. These goals may be irreconcilable in that none are inherently means to attaining the others. In some situations they can be harmonized, but in others they remain in conflict. Such goals and the knowledge surrounding them result from a cultural-historical situation. But in the social goal situation the nuances of a specific cultural-historical situation and the cultural-historical processes for generating such goals have been abstracted away.

Focusing on exactly three concrete goal types may seem overly specific. Indeed, we may not have the goal types quite right. But at some point, we suspect, to be social is to be engaged in highly characteristic dilemmas—much more specific, for instance, than just to have multiple irreconcilable goals. Moreover, these three types of goals have been identified in other research as being the types that are peculiar to the social agent. Research on social dilemma's clearly brings into focus the presence, and indeed potential conflict between self and social goals, and by implication task-related goals (Rapoport, 1961). Weber's (1978, p. 23) typology of ways in which action can be oriented can be mapped onto this typology of goals: instrumental-rational (task), value-rational (individual or social), affectual (individual), and traditional (social). These goals also can be seen as occurring in various ways in Turner's (1988, ch. 5) model of motivation, for example Turner's "need for facticity" is a task-level goal, the "need to sustain self-conception" and the "need for symbolic/material gratification" are individual-level goals, and the "need for sense of group inclusion" is a social-level goal. Thus the agent has goals at the three different levels (task, individual, and social) that may compete and conflict with each other, as well as multiple goals at the same level. Organizational studies of power and institutions often focus on the interplay among these three levels of goals.

### Social Structural Situation (SSS)

Abstracting away these goals results in the social structural situation. This situation is characterized by the presence of groups and group related knowledge such as rules for membership, maintenance, and dissolution. The existence of groups and knowledge about particular classifications of people, such as are embodied in their physiological, demographic, or social-position characteristics (e.g., sex, age, centrality) enable agents to analyze other's actions en mass. An agent in a socially structured situation may identify itself and others by group membership and is constrained in what it knows by its position in the social structure (whom it knows and interacts with).

At this point let us consider what it means to move both right and left along the knowledge dimension. Moving to the right increases the actions available to the agent. For example, knowledge of social structure increases the ability of the agent to plan, over and above simply knowing there are other agents. Situations are listed in a cumulative fashion. For example, the existent of social goals requires a social structure and abstracting away the social structure also abstracts away the social goals and the cultural-historical situation. Agents in less elaborated situations have only a single goal or multiple goals that are reconcilable because they form a means-ends hierarchy. The goals in these earlier situations are all basically defined in terms of self. The existence of social structure permits the separation of task and self; i.e., the agent may now be faced with goals that are demanded of it by an external entity, such as a group or organization. The agent, situated in a social structure, may now be faced with multiple goals that may be irreconcilable by any form of means-ends-analysis as some of these goals are demanded by self, some by task, and some by some external group (the society at large or the organization).

### Real-Time Interactive Situations (RIS)

Abstracting away the social structure leaves the real-time interactive situation. Interactions among agents can actually occur. Other agents, not under the control of the given agent, react to its actions and produce responses that are perceived by the given agent. Such interactions affect the process of cognition itself. Responses must be made within time limits set by the demands of the interactive situation. The agent not only must spend less time interpreting the other agent's action and pondering its own, but must manage its own mental resources and consider the benefits and losses of thinking versus acting. To be struck dumb when asked an embarrassing question often has consequences in and of itself. The requirement of real-time interactive knowledge yields an agent who is aware of the environment and can respond to it in a timely fashion. Since interactions actually occur, there exists a difference between expectations and actuality, and between fantasy and reality.

### Multiple Agent Situations (MAS)

Removing real-time interaction constraints still leaves a multiple agent situation. From a social point of view, the most basic property of an environment is the existence of other agents. The agent must not only deal with these other agents as physical (possibly dynamic) objects, as it would if the object was in some nonsocial task situation. These other agents also must be treated as having goals and taking actions to attain those goals. Depending on what processing model is assumed, the agent must deal with them as knowing everything about the environment that the agent knows, or as having some limited body of knowledge, or as engaging in information processing to determine their actions, or as having emotional reactions to the agent's own actions. The agent must treat them as treating itself and its actions as those of an agent like itself, and so on through the familiar unending reflexive and recursive possibilities. The agent in a multi-agent situation has the knowledge to carry out such reasoning. An agent in this type of situation could acquire "social" goals in that those goals evolve during an interaction and so are the result of cognitive activity but such goals are not social in the sense previously described as they are not the product of a particular cultural-historical environment. This is just the situation envisioned in game theory, although the agents there are often taken to be rational agents.

### Nonsocial Task Situations (NTS)

Removing other agents leaves the task. The nonsocial task situation is devoid of social content. Associated with the nonsocial task situation is knowledge of the task environment and how to attain the task. In order to make specific predictions about the agent's behavior we must provide it with this knowledge. This knowledge however is devoid of social content. Further, the agent comes to a task situation with no knowledge of the social environment that produced the situation or its own social-cultural-historical position. The agent treats other social agents the same way it treats other physical inanimate objects.

To illustrate this point, consider a theorem prover for some logical calculus. As agent, it knows only that its input will arrive as a collection of theorems to be

assumed and a theorem to prove. These may be about anything at all. The theorem prover has a collection of methods to apply, but it uses the same methods on every task. It may, of course, examine the task to see what methods might work, but that is part of its behavior in the situation (and such examinations are part of its repertoire of general methods). The system is quite general, but it has no knowledge at all about the situation (save the form of its input). We may take this as a pardigm of an agent operating in a nonsocial task situation. All social factors—the fact that there are multiple agents, their relative social standing, the history of the theorem it is trying to prove, and so forth, are all irrelevant and not part of the theorem power's repertoire of knowledge. Though the theorem prover can function, prove theorems, and may even make mistakes, we would not claim it to be a social agent.

Next consider Soar, which is the basis for the cognitive agent we defined. It is different from the theorem prover in that its architecture is like that of the human, rather than simply an applier of rules of inference to logic expressions to derive new true expressions. But it is the same as the theorem prover in that, in facing a new situation, it must be given as input knowledge (problem spaces and operators) that describes how to operate in the new situation. Soar is capable of learning and so might learn something by operating in the present situation. However, that would still not be the same as the preparation an adult social agent brings to a new situation. This is due, in part, to the fact that an adult would have operated in a vast number of situations and retained much of the knowledge acquired. The theorem prover and Soar differ on the capability dimension but exhibit only the potential to differ on the knowledge dimension.

The agent in the nonsocial task situation provides the appropriate terminal point for the sequence of agents formed by abstracting the situation. Starting from the most realistic situation we have increasingly abstracting the situation in which the agent is embedded and so have correspondingly abstracted the type of knowledge to which the agent has access. The more abstract the situation the fewer the possible actions. Thus, abstracting away categories of knowledge decreases the agent's ability to act by narrowing the range of possibilities to just those that are appropriate in that situation.

### Summary

We have laid out both dimensions for defining a candidate Model Social Agent. Figure 1 shows the resultant matrix. Each cell of this matrix defines a particular abstract social agent, the degree of abstraction diminishing as one moves down and to the right. Thus the prime candidate for an actual Model Social Agent is the cell in the lower right-hand corner. It is an emotional-cognitive agent in a particular cultural-historical situation with the possibility of accessing all the attendant knowledge. The Model Social Agent does not necessarily know everything there is to know about the situation, rather, the Model Social Agent is not missing a "class" of knowledge.

One does not need to be at the endpoint on both the information-processing capabilities and the knowledge dimension to explain many behaviors. As each limitation is placed on the agent's capabilities the agent emerges with more, often more complex, behaviors. Similarly, as the situation the agent knows about becomes increasingly less abstract, behaviors become increasingly complex. Figure 3 illustrates

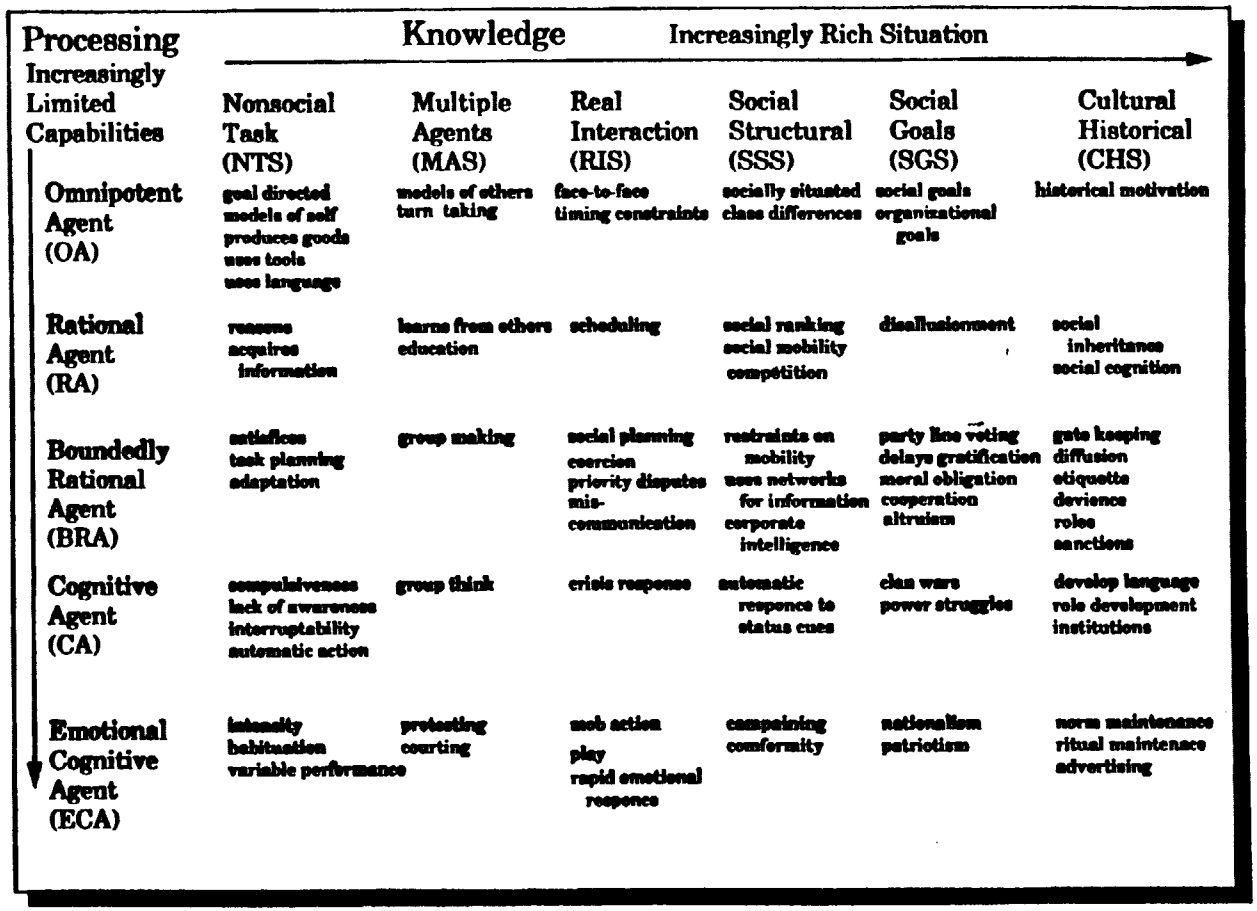


FIGURE 3. Application of fractionation matrix to social behavior.

this by noting for each cell in our framework some typical additional behaviors that an agent with those capabilities and knowledge should be able to exhibit. As you change agents by moving to the right or down, that agent is presumed to be able to exhibit all the behaviors up and to the left, and in addition, the new behaviors listed in that cell. Thus, the MSA in the ECA, CHS cell would seem to be capable of all behaviors shown in the figure. Each behavior is listed by name in the cell that corresponds to the capabilities and content minimally necessary to produce this behavior. This figure is just illustrative, to help ground the analytical scheme. No attempt has been made to be definitive.

When placing behaviors in this fractionation matrix each item was located as far up as possible on the capabilities dimension and as far left as possible on the knowledge dimension. This positioning reflects the objective of establishing each type of social behavior with as few assumptions as possible. Consider party-line voting [BRA, SGS]. This behavior requires at least that there exists a social structure, such as an organization (and in this case a political party), that establishes a social goal (vote for the party) that it then places on its members. Whether or not cultural historical information is needed is unclear. This defines the situation to be a Social Goal Situation. The agent must, of course, have knowledge of this goal in order to restrict its behavior in this fashion, and may have its own alternate goal that it ignores. Such ignorance, arises from a satisficing behavior, where the agent has incomplete information and a lack of time to gather more, and so satisfices by going along with the social goal. Satisficing, or any action that causes the agent to supplant a group goal with its own, requires some type of bound on agent rationality. This defines the agent as being at least Boundedly Rational.

Let us consider other behaviors. First let us consider those behaviors requiring no situational knowledge other than that available in the most generic Nonsocial Task Situation. As we restrict the agent's capability increasingly complex behaviors appear. The Omnipotent Agent (OA) is goal directed, has models of self, can produce goods, uses tools, and uses language. However, as previously stated, the Omnipotent Agent is omniscient within a particular situation. Since the agent is omniscient and knows everything it has no need to reason, or to acquire information. Those capabilities arise from restricting the agent's capabilities, but only slightly. The rational agent must acquire information but it can acquire and use (i.e., reason) all information in the situation. The rational agent because it can acquire all information, does not need to adapt or plan. By placing bounds on the agent's rationality, we get the need for satisficing, adaptation, and planning. Adaptation and planning are responses to the lack of information, or the inability to process information. The Cognitive Agent (CA), unlike the Boundedly Rational Agent (BRA), acts at a micro level, without awareness, and has a specified perception and action system. These further constraints generate compulsive action (action automatically triggered by micro decisions, without the subjects awareness). Further, within the CA cognition is in service of the other systems, thus the agent, unlike the BRA, responds to interruptions form the environment. The CA will, however, operate in a very staid and consistent fashion. Unless the CA acquires new information, it will exhibit identical performance in the identical situation. Further constraining the agent by adding emotions results in variable performance, and modulates the intensity with

which the agent responds to situations, and enables habituation. This is because, in part, emotions vary in the extent to which they affect the salience of different information.

Now let us hold constant the agent's capabilities and vary the situation. Consider the Omnipotent Agent (OA). As we move from the Nonsocial Task Situation (NTS) to the Multiple Agents Situation (MAS) additional agents are added. Consequently, the OA now needs models of others and can engage in turn taking. Such behaviors were not needed in the more simple situation. The Real Interaction Situation (RIS) adds the need to operate in real time and to consider others, not in the abstract, but as physically present interaction partners. This results in the need for behaviors such as face-to-face interaction and attention to temporal constraints such as it takes "$x$" minutes to do this task. In a Social Structural Situation, unlike the previous situations, there exists an extant social structure (that may or may not be governed by its own laws). To cope with this situation the agent must now know that it is socially situated, and that there are class differences. Of course the Omnipotent agent, because it is omniscient within a situation will have a perfect understanding of the social structure and its position in it. Other more restricted agents have an imperfect understanding which then results in a wider range of behaviors relative to the social structure. Moving on, in the Social Goal Situation (SGS) there exist multiple competing goals. The agent in this situation now can take on these goals, and can switch among them in determining its actions. The Cultural Historical Situation (CHS) brings with it added relevance to the underlying culture and the history in which current action is embedded. For the omnipotent agent, who already knows everything, all this adds is the need for being motivated by the past or by this culture.

The position of all other behaviors that appear in Figure 3 can be determined by combining the behavior associated with the OA in the task of interest and the behavior associated with the NTS for the agent of interest. For example, consider group think [CA, MAS]. At a minimum group think involves a certain lack of awareness as well as aligning one's model of reality with the models of others (which requires there to be others). Since group think involves others it cannot occur in NTS. Since group think involves awareness it cannot occur for the BRA. Thus the minimal agent and situation which appears necessary to produce it is CA, MAS. Similarly, competition [RA, SSS] requires there to exist at least a situation with an extant social situation and an agent who at a minimum does not know everything and must try to acquire information. In contrast, ritual maintenance [ECA, CHS] requires knowledge of the underlying culture and history, and the existence of ritual which implies that the agent is in CHS. To maintain a ritual, however, agents must rely on intensity of response, and the ability of agents to habituate actions, which implies that the agent is at least ECA. Similar arguments underlie the placement of the remaining behaviors.

## APPLYING THE FRACTIONATION MATRIX

The fractionation matrix has conceptual and analytical uses in addition to providing guidelines for the construction of operational Model Social Agents. Since the

scheme is not monolithic, but fractionates the characteristics involved into a whole series of capabilities and enabling situations, it can be used to shed light on the models of the social agent that are used in current social sciences. This is not the main theoretical use one hopes for from the fractionation matrix, which is to place the model within some larger social-theory enterprise. But using it to analyze current work will permit some assessment of the fruitfulness of the proposed fractionation matrix.

We try our hand at three examples. The first is an analysis of a socially relevant theory about the individual, Festinger's Social Comparison Theory. The second is an analysis of a sweeping theory of social structure and process, Turner's Social Interaction Theory. The third is an attempt to make some sense of the many different views of humans that have been put forward. Each analysis is very brief and hardly definitive—each is worthy of an entire paper all by itself. Thus, these are really just exercises to illustrate the potential of the analytical scheme.

### Festinger's Social Comparison Theory

As our first exercise, we address the extent to which a familiar piece of social psychology, *Social Comparison Theory* (*SCT*), set forth 40 years ago by Leon Festinger (1954), embodies a model of the social agent.[14] SCT addresses what sort of decisions a person makes in a social situation. It posits that the person bases his or her decision on a comparison with analogous decisions made by other persons. It remains a live theory, which continues to be used in analyzing many social situations and continues to be developed theoretically (Suls and Miller, 1977). The question we wish to address is what role is played by the different aspects of the social agent we have teased out. We begin by noting that this theory when placed within the context of the fractionation matrix would lie in the cell (emotional-cognitive agent, multiple-agent situation). As such, SCT cannot be thought of as a full model of the social agent. The question that comes to the fore is does having a fractionation matrix help; i.e., are we doing anything more than simply placing SCT on a grid?

A nice thing about Festinger's style of theorizing is his tendency to write down an explicit set of axioms. Thus, we can classify the axioms for SCT with respect to the type of agent being posited—which axioms refer simply to the agent being an omnipotent agent [OA], which to the rational agent [RA], and so on—and by the type of situation being posited—which axioms refer to the task [NTS], which to the multi-agent situation [MAS], and so on. Table 4 lists the hypotheses, suitably classified. It is not necessary to go through the hypotheses in detail. But consider a couple, just for illustration. Hypothesis 1 says there exists a drive for everyone to evaluate. That is certainly a basic tenet of general cognitive behavior in support of rational behavior. The CA evaluates all the time. It does it under the impress of a goal to be attained, whereas SCT ascribes it to a basic drive. But a review of the way Festinger uses the theory shows that this is a distinction without much difference—the effect of hypothesis 1 is to permit the analyst to posit an evaluation anywhere he finds it expedient. Further, hypothesis 1 is a statement only about the agent and

[14] This analysis is an expanded version of an analysis in Newell (1990).

TABLE 4
Tenets of Social Comparison Theory (Festinger, 1954)

| | |
|---|---|
| 1. | "There exists, in the human organism, a drive to evaluate his opinions and his abilities." [CA, NS] |
| 2. | "To the extent that objective, non-social means are not available, people evaluate their opinions and abilities by comparison respectively with the opinions and abilities of others." [CA, MAS] |
| 3. | "The tendency to compare oneself with some other specific person decreases as the difference between his opinion or ability and one's own increases." [CA, MAS] |
| 4. | "There is an unidirectional drive upward in the case of abilities which is largely absent in opinions." [CA, NTS] |
| 5. | "There are non-social restraints which make it difficult or even impossible to change one's ability. These non-social restraints are largely absent for opinions." [CA, NTS] |
| 6. | "The cessation of comparison with others is accompanied by hostility or derogation to the extent that continued comparison with those persons implies unpleasant consequences." [ECA, MAS] |
| 7. | "Any factors which increase the importance of some particular group as a comparison group for some particular opinion or ability will increase the pressure toward uniformity concerning that ability or opinion within that group." [CA, SSS] |
| 8. | "If persons who are very divergent from one's own opinion or ability are perceived as different from oneself on attributes consistent with the divergence, the tendency to narrow the range of comparability becomes stronger." [CA, MAS] |
| 9. | "When there is a range of opinion or ability in a group, the relative strength of the three manifestations of pressures toward uniformity will be different for those who are close to the mode of the group than for those who are distant from the mode ... " [CA, SSS] |

there are no specifications on the situation. Thus, we classify hypothesis 1 (as well as 4 and 5 for analogous reasons) as [CA, NTS].

Hypothesis 2 specifies that if objective evaluation is not possible, then evaluation proceeds through comparison with others. This is essentially a principle of cognitive operation in a multi-agent situation. If evaluation is necessary then other's evaluations are a source of relevant knowledge. Thus, this hypothesis (as well as 3 and 8) is classified as [CA, MAS]. There is no specific social or cultural content involved and no sense of group qua group. We note that even though Festinger uses the nomenclature of groups, most of the time he is really talking about dyadic exchange and the mere fact that multiple agents are present and not special group-level knowledge. In contrast, hypothesis 7 specifies the existence of groups as entities with varying importance to the individual and hypothesis 9 specifies that the position of the individual in the group (i.e., the individual's structural position) affects behavior. These two hypotheses (7 and 9) with their emphasis on groups qua groups, and not just a collection of independent agents, can be classified as [CA, SSS]. Of these two hypotheses, 9 is particularly interesting, as it specifies particular knowledge that the agent has as a group member, knowledge that is peculiar to position within the group.

Each of the hypotheses can be classified in this way. The result is that only one, hypothesis 6, is not CA. This is the hypothesis that hostility and derogation will accompany the act of ceasing to compare. If an agent has been using someone as the basis of comparison, and then ceases to do so, then the agent dumps on the other person. That does not follow from any obvious rational considerations, even in the multi-agent situation, and it is clearly emotional behavior, so it is labeled ECA. Although the SCT agent is "cognition + emotion", the notion of emotional response is quite underdeveloped. For example, SCT relies only on a single dimension of emotional behavior (positive and negative), whereas many studies have indicated the presence of multiple dimensions (Ortony, Clore and Foss, 1987; Heise, 1977; Heise, 1978; Heise, 1979; Kemper, 1987). A consequence of this lack of development is that a wide range of social behavior, largely those related to emotional reaction, do not arise in the SCT agent. As an example, the SCT agent's language comprehension is not affected by emotion, as is human children's (Ridgeway, 1985) and the cognitive capabilities, action tendencies, and physiological activity of the SCT agent, unlike the full ECA agent, are not comprised or enhanced by its emotional state (Ax, 1953; Frijda, 1987).

The interesting conclusion from this brief analysis is that, to get some theoretical action in the domain of social comparison, Festinger had to provide an entire structure that was mostly nonsocial. In keeping with the analysis, we have laid out the tenets of social comparison theory, and the work done by Festinger in this area, as represented in the article (Festinger, 1954) relative to the fractionation matrix that we have put forward (see Figure 4). Each of the hypotheses (H), corollaries (C), and derivations (D) are listed by the number under which they occur in (Festinger, 1954) in the cell which characterizes the agent/situation that they presume. In addition, we have taken 9 other papers by Festinger and colleagues that are associated with SCT and have placed these by author in the appropriate cell. It is readily apparent that the bulk of theory is in the cell—cognitive agent, multi-agent situation, appropriate cell. Figure 4 graphically illustrates that Festinger had to spend most of his theoretical degrees of freedom building up the basic apparatus of choice behavior. The elements of a social situation that are necessary for SCT are primarily the existence of other social actors, not any strong posits about the contents of their beliefs. Many of the inferences of SCT do not follow from strongly social or cultural aspects. Mostly, they follow just from the objective character of the task environment, such as the reliability of decision criteria, and the mere extension to the existence of other rational agents. Thus, most of the knowledge possessed by the SCT agent is task knowledge. For example, let us contrast the expected behavior of a ballerina in a national company and a third grader taking ballet classes who are both asked whether they are good dancers. According to SCT they will locate their referent group (professional ballerinas versus other third graders) and then will respond (adequate and very good respectively) based on these groups. Making this judgment, requires knowledge of the task (dancing), the ability to distinguish a referent group, and knowledge of the group members' abilities. In other words, task knowledge and general cognition are all that is required.

To the extent that Festinger did have a social theory in SCT, it centered on the assumption that there was an entity called the group that individuals could reason

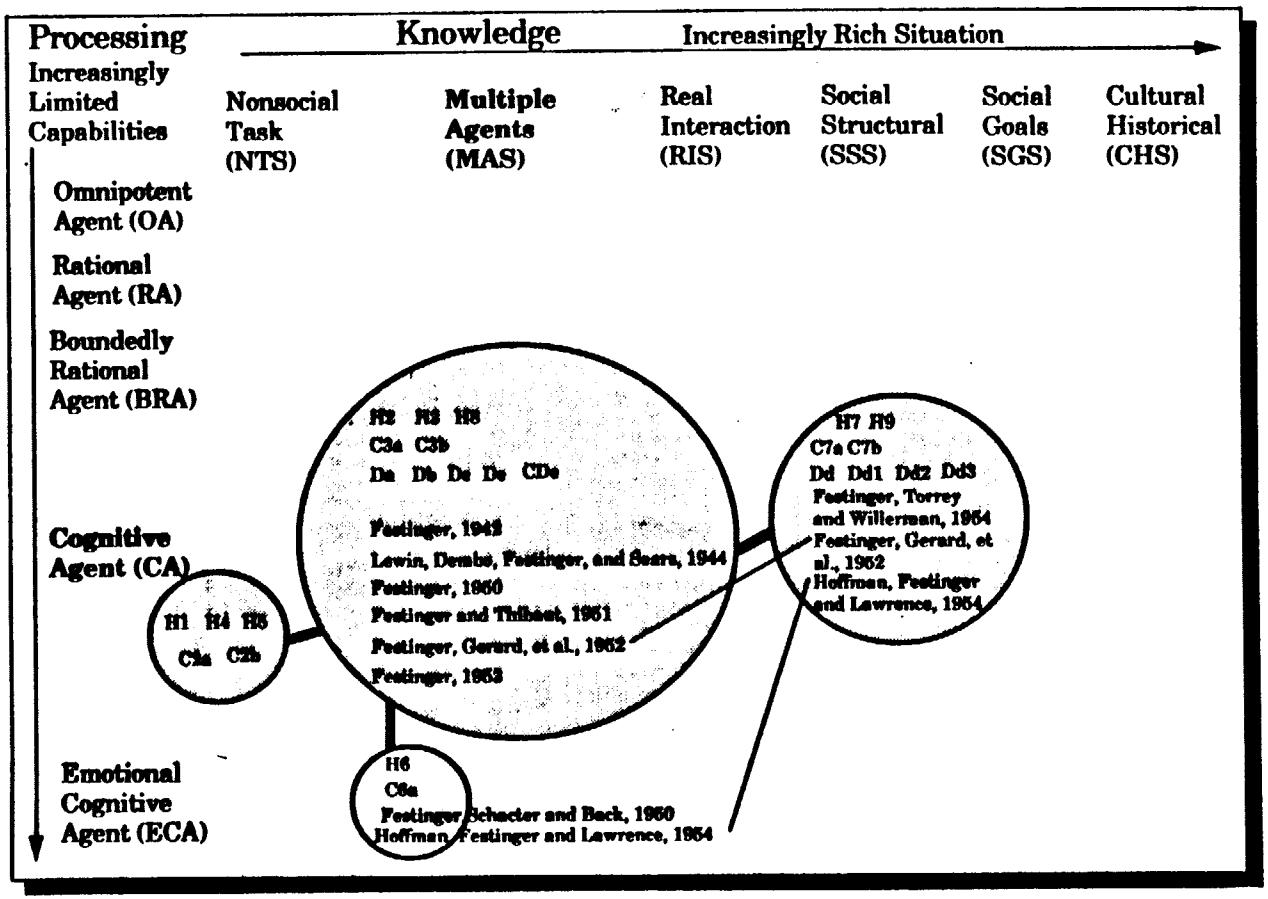


FIGURE 4. Festinger's social comparison theory.

about and that individuals possessed knowledge about their and other's relative position in the group so as to act on the basis of their relative position. Here, however, the theory is very underspecified. Recall that relative to the knowledge dimension, there exists an additional dimension of *specificity*. If the theory does not provide specific details, these can be defined arbitrarily and the theory should still hold. Thus, according to hypothesis 7, regardless of how the group is defined the pressure toward uniformity should increase as the importance of the group as a comparison group grows.

Possibly, with a general theory of the social agent, such as that described in the present paper, more substantial progress could occur in the enterprise of which SCT is one part. We expect that by finally getting an appropriate theory of the individual agent at the cognitive level (such as CA)—and not just at the knowledge level, which is what game theory and its psychological derivative, decision theory, basically provide—this can become common ground for all such efforts. The pattern of pervasive limits to rationality revealed by behavioral decision theory already indicates something of the complexity of the total picture (Kahneman, Slovic and Tversky, 1982; Pitz and Sachs, 1984), though it does not integrate these results into a unified architecture. We further expect that it will become possible to generate a model of the agent in the social structural situation that goes beyond claims that recognizing groups and position in groups affect behavior. In particular, we expect that given a collection of such agents it will be possible to examine the development of such structural phenomena as socially shared cognitions and distributed problem solving. In fact, we expect that given a collection of such agents it will be possible to address the extent to which social reality is a completely emergent phenomena.[15]

### Turner's Social Interaction Theory

As our second exercise, we address the extent to which a recent general theory, *Social Interaction Theory* (*SIT*) as set forth by Jonathan Turner (1988), embodies a model of the social agent. SIT is concerned with providing an integrated theory of interaction, which has motivation as its driving function and which encompasses the development of different forms of interaction and the social/cultural consequences that derive from these. As in the previous section, we wish to address what role is played by the different aspects of the social agent we have teased out.

Unlike Festinger, Turner does not provide us with a completely axiomatized model. Rather, Turner provides an analysis at two levels. At the high level he diagrams the relationship among variables and then, at the second level, he asserts a series of propositions about some of the entities in the diagrams. Unlike Festinger, these propositions are not axioms; i.e., other propositions can be derived from the diagram. A total of 30 propositions are listed in the book. The main propositions are divided, as is the theory, into three areas—motivation (3 propositions, ch. 5), interaction (9 propositions, ch. 8), and structuring (6 propositions, ch. 11). These 18 propositions are listed in Table 5. We identify the section in which the proposition

[15]The work by Carley et al. (1992) on Plural Soar is an attempt at determining what social behavior emerges from the ongoing interactions of a collection of Soar agents.

TABLE 5
Propositions of Social Interaction Theory (Turner, 1988)

MOTIVATION PROPOSITIONS

M1 p. 67 "The overall level of motivational energy of an individual during interaction is a steep *s*-function of the level of diffuse anxiety experienced by that individual." [ERA, MAS]

M2 p. 67 "The overall level of diffuse anxiety of an individual during interaction is an inverse and additive function of the extent to which needs for group inclusion, trust, ontological security, and confirmation/affirmation of self are being met." [ERA, MAS]

M3 p. 68 "The degree to which an individual will seek to maintain an interaction, or to renew and reproduce it at subsequent points in time, is an additive function of the extent to which needs for group inclusion, trust, ontological security, self confirmation/affirmation, gratification, and facticity are being met." [RA, MAS]

INTERACTION PROPOSITIONS

I1 p. 116 "The degree of interaction between two or more actors is an additive function of their level of signaling and interpreting." [RA, MAS]

I2 p. 116 "The degree of interaction between two or more actors is an additive function of their level of signaling and interpreting." [RA, MAS]

I3 p. 116-7 "The level of role-taking in an interaction is a primary function of the degree of visibility in the ritual-making and stage-making gestures of others and a secondary function of the level of ability in using stocks of knowledge to understand the frame-making gestures of others." [RA, CHS]

I4 p. 117 "The level of frame-making in an interaction is a primary function of the level of ability in using appropriate stocks of knowledge to make claims and construct accounts and a secondary function of the degree of intensity in role-making." [RA, CHS]

I5 p. 117 "The level of frame-taking in an interaction is a primary function of the degree of visibility in the claim-making and claim-taking gestures of others, and a secondary function of the degree of intensity in role-taking with others." [RA, CHS]

I6 p. 117 "The level of stage-making/taking in an interaction is an additive function of the level of role-making/taking and ritual-making/taking." [RA, CHS]

I7 p. 117 "The level of ritual-making/taking in an interaction is an additive function of the level of role-making/taking and stage-making/taking." [RA, CHS]

I8 p. 117 "The level of account-making/taking in an interaction is an additive function of the level of claim-making/taking and frame-making/taking." [RA, CHS]

I9 p. 117 "The level of claim-making/taking in an interaction is an additive function of the level of account-making/taking and frame-making/taking." [RA, CHS]

STRUCTURATION PROPOSITIONS

S1 p. 172 "The degree to which individuals reveal consensual agreements about the level of intimacy, ceremony, and socializing required in a situation (categorize) is a positive and additive function of the extent to which they regionalize and normatize." [RA, CHS]

TABLE 5
*(Continued)*

| | |
|---|---|
| S2 p. 172 | "The degree to which individuals share knowledge about the meaning of the objects, physical divisions, and distributions of people in space (regionalize) is a positive and additive function of the extent to which they routinize and categorize." [RA, SSS] |
| S3 p. 172 | "The degree to which individuals construct agreements about the rights, duties, and interpersonal schemata appropriate to a situation (normatize) is a positive and additive function of the extent to which they categorize, ritualize, and stabilize resource transfers." [RA, CHS] |
| S4 p. 172 | "The degree to which individuals agree upon the opening, closing, forming, totemizing, and repairing behavioral sequences relevant to a situation (ritualize) is a positive and additive function of the degree to which they regionalize, categorize, normatize, and stabilize resource transfers." [RA, CHS] |
| S5 p. 172 | "The degree to which individuals develop compatible as well as habitual behavioral and interpersonal responses to a situation (routinize) is a positive and additive function of the extent to which they regionalize and stabilize resource transfers." [RA, CHS] |
| S6 p. 172 | "The degree to which individuals accept as appropriate a given ratio of resource transfers in a situation (stabilize) is a positive and additive function of the extent to which they normatize, ritualize, and routinize." [RA, CHS] |

appeared by using the letters—M, I, and S. In addition to the 18 main propositions, Turner lists an additional 12 propositions in his concluding chapter.

We classify each of these propositions by the type of agent being postulated (Table 5) and then place the symbol for that proposition in the appropriate cell in Figure 5. The tenets of social interaction theory as described by Turner (Turner, 1988) are laid out relative to the fractionation matrix that we have put forward. Each of the key propositions in Chapters 5, 8, 11, and 13 are listed by the number under which they occur in (Turner, 1988) in the cell that characterizes the agent/situation that they presume. We append a letter to the front of these numbers in order to indicate the section they are from in Turner's book. Thus M are the motivation propositions (ch. 5), I are the interaction propositions (ch. 8), S are the structuring propositions (ch. 11), IS are the impact of self propositions (ch. 13), IFI are the impact of feeling involved propositions (ch. 13), and IFR are the impact of feeling right propositions (ch. 13). In addition, we have taken the various categorization and cognization schemes laid out in these chapters and mapped them onto our framework. It is readily apparent that the bulk of theory is in the cell, [CA, CHS].

For example, consider proposition I1 that *"the degree of interaction between two or more actors is an additive function of their level of signaling and interpreting"* (Turner, 1988). Since there are multiple actors, the situation is at least a multi-agent situation. Since Turner never specifies in the discussion any real-time requirements, the situation is at most multi-agent. Signaling simply involves the communication of information and does not place any processing limitations on the agent. Interpretation, at first appears to place processing constraints on the agent. After all, interpretation seems to imply that the agent has limiting processing capabilities and

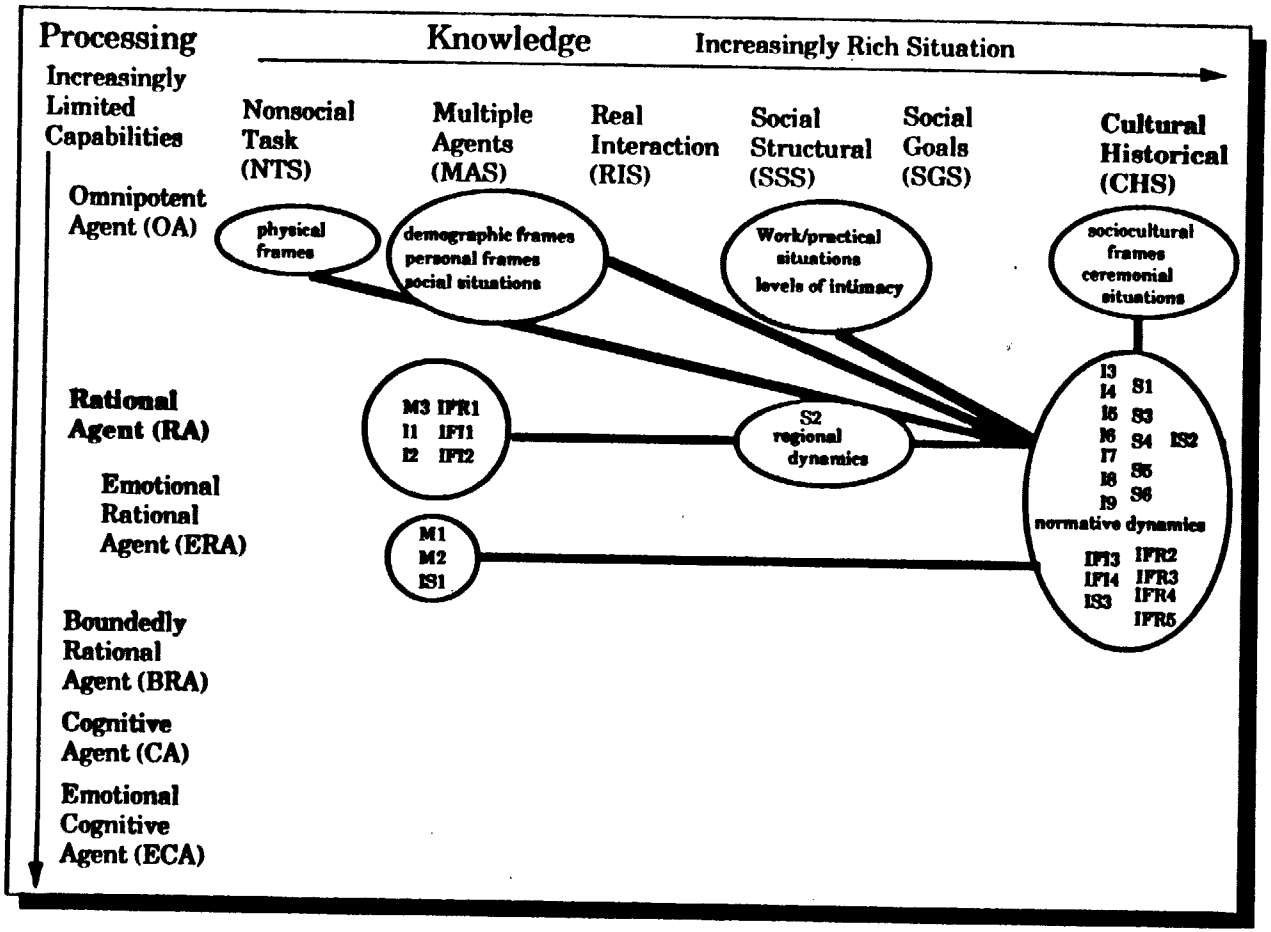


FIGURE 5. Turner's social interaction theory.

so cannot make use of all knowledge it has at its disposal. However, a detailed reading of chapter 9 reveals that Turner's agent is a purely knowledge-level; i.e. rational, agent. There is no mention of processing constraints only knowledge constraints. Interpretation is a function of available knowledge. We follow this same procedure for all 30 propositions. In addition to classifying all 30 propositions, also we classify the dominant variables (such as types of frames) identified by Turner throughout the book.

Key propositions from Turner's theory of motivation during interaction (ch. 5) are by necessity in the multi-agent situation, because Turner is concerned with interactions among agents. In these propositions, we note that Turner, like Festinger, uses an emotion (anxiety/hostility) as part of the mechanism for regulating behavior. In contrast, most of the key propositions regarding interaction or structuration require a cultural-historical situation. In these propositions Turner is not specifying a specific environment but is specifying meta-knowledge that an agent in such an environment would have and the method by which the agent will use cultural-historical knowledge.

The interesting conclusion from this analysis is that Turner is relying on a very powerful agent; i.e., the rational agent in a cultural-historical situation. As is illustrated graphically in Figure 5, Turner spends most of his theoretical degrees of freedom building up meta-knowledge in the cultural-historical situation which only limits the agent in terms of what type of knowledge it has. If actually implemented, Turner's social agent would exhibit pure rationality and would be able to rapidly process vast quantities of information with no side effects.

### The Many Models of Man

As our third and final exercise, we come back to the plethora of views of humans extant in the social sciences, which we noted at the very beginning of the paper. We do this by positioning many theories of social agents in the space defined by our framework (Figure 6). Theories are listed by name int he cell which characterizes their predominant assumption. That is, theories are placed in the cell that most nearly characterizes the major assumptiosn in that theory. We do try to be generous; i.e., we place theories as far right and down as they seem to warrant. Unlike the analyses of Festinger (SCT) and Turner (SIT), where we could show the way a single theory distributes itself over many types of models, here each theory must be indicated by a single point (and both SCT and SIT can be found in Figure 6). For instance, in terms of knowledge, the theories have been placed at the dominant type of knowledge posited by that perspective; however, these theories generally do not take explicit account of all of the types of knowledge to their left. Most of the theories are purely descriptive. Thus, the social agents they presuppose are not specified concretely enough to exist as operational agents. We list those that exist in an operational form in bold face type.

This exercise serves a variety of purposes. It shows the generality of the fractionation matrix; i.e., we are able to use it to make general sense of a large number of theories, not just Festinger's social comparison theory and Turner's social interaction theory. It shows the limits of current theories of the social agent. Specifically, by comparing Figures 3 and 6, one can see the social behaviors that a theory of

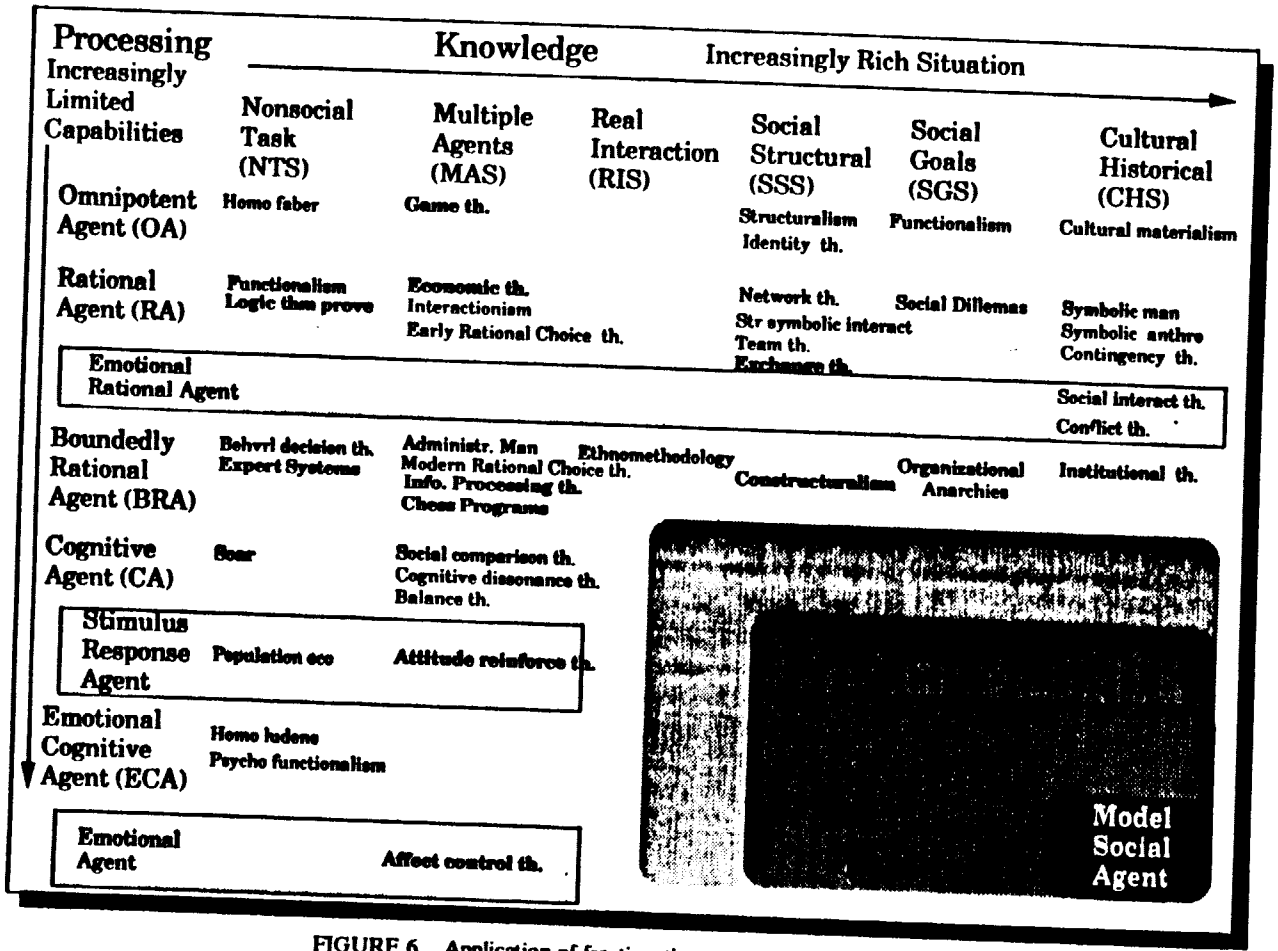


FIGURE 6. Application of fractionation matrix to social theories.

social agent with particular capabilities and content can be expected to explain. A theory in a particular cell (Figure 6) cannot be expected to fully explain the behaviors (Figure 3) in cells to the right and down from its position. For example, Turner's social interaction theory [RA, CHS] should be able to explain disallusionment and altruism but not the use of social networks for information gathering. Carley's constructuralism [BRA, SSS] (Carley, 1991; Carley, 1990), which is highly similar to Turner's theory except that it relies on agents who are at least boundedly rational and do not have irreconcilable goals, should have the opposite theoretical limitations. Neither, theory should be able to offer a complete explanation of behaviors such as ritual maintenance.

In placing these theories we went through exercises like that portrayed for Turner and Festinger for each theory. Let us consider two more of these theories. Classical game theory requires two players who interact but not in real time or with any knowledge of the social structure. They are thus in the Multiple Agents Situation (MAS). Further, in the classical formulations (as opposed to current variants) the agents are completely unrestricted (OA); i.e., they have complete knowledge, foresight, no restrictions on their actions, etc. Whereas, institutional theory [BRA, CHS] assumes agents that not only have incomplete information but that satisfice (hence BRA) as they learn, track, and respond to the ongoing performance of other organizations in the environment and can adapt so that they take on the form of these other organizations (and therefore must have cultural and historical knowledge).

The analysis in Figure 6 illustrates that social studies of human nature have indeed been cumulative. Modern organizational theories, for example, have moved well beyond *homo faber*—the omnipotent agent in the nonsocial task situation. But, this analysis also illustrates where there are gaps in the extant models of the social agent and thereby provides direction for how current models should be extended. Specifically, we see that there is not a single theory that combines a strong processing model with a sufficiently complex knowledge-based environment.

In doing this exercise we found only two additional positions on the processing dimension—the *stimulus-response agent* [SRA] and the purely *emotional agent* [EA]. The development of cognitive science has shown that the stimulus-response agent is much too limited a set of mechanisms to actually form the basis of an operational agent. But historically it has played a large role, so it is the sort of processing agent presumed by many efforts in the social sciences. One can try retroactively to stretch the stimulus-response agent to give it additional cognitive powers, as Turner does (Turner, 1988, p. 27): " ... even extreme behaviorism implicitly invoked cognitive processes, since it assumed that responses that brought gratification would be retained." However, behaviorism tends to ignore that which is distinctly human—complex, but limited cognitive capacities. The same argument can be made of the purely emotional agent.

The Festinger and Turner analyses in the previous sections serve to illustrate how social theories fit within this framework. Let us consider two others briefly, Huizinga's *homo ludens* or playful human and institutional theory. As noted by Huizinga (1950), "There is a third function, however, applicable to both human and animal life, and just as important as reasoning and making—namely,

playing." We have placed this theory in the cell emotional-cognitive agent and multiple agent situation. Huizinga's conception clearly has additional limitations that go beyond those present in the cognitive agent. Playful human exhibits emotions and needs emotions to respond to the world. Playful human also requires at least a multi-agent situation as Huizinga constantly alludes to the presence of others. A real interaction situation is not necessary as play to be play need not physically occur but can simply be thought about or imagined. Detailed review shows no reliance by Huizinga on any more complex features of the environment. Thus *homo ludens* is placed in this cell [ECA, NTS] as this is the type of agent Huizinga relies on. By placing *homo ludens* in this position the claim is being made that all of the features of play defined by Huizinga (e.g. creativity, imagination, and pretend interaction) can be exhibited by a model agent with these capabilities. This analysis, of course, is of Huizinga's theory of play; it does not reach out to consider whether this theory is actually adequate to describe playful behavior of humans. Similarly, all entries in Figure 6 characterize the agent assumed by the theories, but they do not assess how adequate are those agents to the empirical phenomena.

Consider, for example, institutional theory which argues that organizations imitate each other in order to minimize sanctions from stakeholders. Such a statement is assuming implicitly that there are multiple agents, taking actions in real time, with positions (as stakeholders) within a social structure, having organizational goals, and capable of taking action as a group (sanctions) that are culturally and historically determined. Moreover, the mere fact that imitation can occur implies that diffusion has taken place. Thus the organizational agents of institutional theory are, at least implicitly boundedly rational agents in cultural-historical situations [BRA, CHS].

## DISCUSSION

There are other advantages to approaching the nature of the social agent through a model that can be realized as an artificial agent. In analogy to the classic Turing Test (Turing, 1950), it allows us to imagine an ideal operational test of a set of hypotheses about the social agent:

> **The Social Turing Test:** Construct a collection of artificial social agents according to the hypotheses about what makes agents social and put them in a social situation, as defined by the hypotheses. Then recognizably social behavior should result. Aspects not specified by the hypotheses, of which there will be many, can be determined at will. The behavior of the system can vary widely with such specification, but it should remain recognizably social.

The Social Turing Test, like the Turing Test, is a sufficiency test that depends on human recognition. It tests the proposition, *if the agent has properties x, y and z, then it behaves socially*, on the assumption that humans can recognize social behavior in all of its forms. The Social Turing Test is both weaker and stronger than the Turing Test. It is weaker because it does not require confusing a computer with a person. It is stronger because one can plug in many values for those aspects not specified. Although actually carrying out such a test is well beyond the current art, it is useful to keep in mind, since it expresses clearly in what image a model of the social agent should be constructed and criticized. Although much weaker than a genuine theory of the social agent, it can serve as a heuristic guide.

We have arrived at a candidate Model Social Agent. We certainly do not know whether it would be adequate yet to pass the stringent test we put forth at the beginning, namely, that a collection of such agents, having these properties guaranteed and no others, put in a social situation (as stipulated), would produce recognizable social behavior. Such a test requires synthetic agents that are well beyond the simulation art. We do not even know how a collection of agents defined at much less elaborate cells in the process-knowledge matrix would behave—would they seem social in some interesting ways or present a caricature of social behavior? All we know is that many important aspects of being human and social are included, even though some of them, in particular emotion, have had to remain sketchy.

Actual artificial social agents would lend themselves to many interesting simulation uses besides the Social Turing Test, though most are still beyond reach. Still, the *Model Social Agent*, as an explicit model of the human to be used within the social domain remains a useful conceptual move. The Model Social Agent operates as a repository of essential properties. By being a definite abstraction, it permits definitive analysis, so that arguments always do not have to appeal to an indefinitely rich but unarticulated notion of humanity. It lets everyone use the *same* notion of the human, thus helping to factor cleanly what are issues of sociology and organizations from what are issues of psychology. It abets the accumulation of important implications of being human for social systems. As these implications are used to enrich the model of the social agent itself they become uniformly availability for analysis. For example, as the role of interruption in human social action becomes clear, mechanisms for handling interruption become built into the model. This makes that characteristic manifest to all users of the model. In enumerating these good things, it is not our purpose to oversell the notion. It is, after all, only a conceptual device. It will evoke its share of scholarly disagreement and be characterizable only within a certain margin of fuzziness.

As a final point, when defining a model agent there is in essence another dimension within the knowledge dimension. We refer to this additional dimension as *specificity*. Specificity refers to the level of information available given a specific level of abstraction (situation). It is useful to distinguish four levels[16]—existence, quantity, structure, and content. *Existence* means that the model specifies only that there exists knowledge of the given category, but nothing about what that knowledge is. *Quantity* means that the model specifies the amount of knowledge. *Structure* means that the model specifies how the agent's knowledge lies within some network or hierarchy of categories. *Content* means that the model specifies what would actually be known by a human agent in the situation. The specificity of the knowledge determines what predictions can be made by the model. Much less can be predicted with a model if it posits of an agent that it "values things" (an existence statement) rather than that it "values socializing with others" (a structure statement) rather than it "enjoys going to dinner with Jane" (a content statement). As an aside, we note that a great deal of theorizing within the social and organizational sciences, and many of the empirical tests, center on determining which of several structure

[16]Additional levels are possible; these four seem to cover most models in sociology and organizational science.

statements are correct (e.g., do people value socializing with similar others or do people value socializing with dissimilar others).

Normally, sociological and organizational models describe the human agent at low levels of specification. Lack of specificity shows up as the inability to make specific predictions, as in the example above. On the other hand a candidate Model Social Agent for the Social Turing Test must be operational. The agent must be able to act and respond in actual situations, and the agent must have knowledge of some sort to enable it to do so. The agent cannot just "value things"; the agent cannot even just "value socializing with others"; the agent must have, or be capable of generating, some specific knowledge, such as "value socializing with Joe". In line with the Social Turing Test, less specificity than full content can be taken to mean that the additional specification can be provided arbitrarily.

## HOW CLOSE ARE WE TO THE ARTIFICIAL SOCIAL AGENT

It is instructive to contrast the conception of the social agent put forward in this paper (see Figure 1) with Soar, the operational basis for our cognitive agent. Recall our earlier discussion of the theorem prover and Soar. We noted that they differed in capability but had only the potential to differ in knowledge. The theorem prover and Soar are quite general in scope but they lack the knowledge they should have as they enter a new situation. The theorem prover and Soar are not inherently debarred from having such knowledge. The theorem prover could have a large set of theorems available to it as it enters the room and Soar also could have a large collection of symbol structures. It is doubtful that they could have *all* the knowledge they would ever use as, after all, each new situation brings with it new information. But then, the social agent need not have all the knowledge, just the right types of knowledge. The theorem prover is even more domain specific than Soar as the only kind of task it can work on are theorem proofs. The theorem prover only has task goals (as opposed to having social goals). The goal always is to solve the task regardless of the situation generating the task. In contrast, the social agent often is subjugated to social goals. Soar is closer to the social agent in that it has a full architecture and can handle (given the right knowledge) multiple tasks and have multiple goals. Soar has generally been applied to nonsocial situations, to tasks such as chess or theorem proving, where social goals are irrelevant. In one case, Soar was applied to a social situation—an organization of agents working collectively to fill orders in a warehouse (Carley, et al., 1992). Yet even in this case, the task, individual, and social goal were reconcilable. Whether Soar can accomplish social goals that are irreconcilable with task and individual goals is an open question.

To convert Soar, which is an actual working artificial cognitive agent, into a social agent one must do more than simply add the right knowledge. Imagine that we were to treat Soar as an infant and socialize it by giving it the entire rich body of knowledge we were previously describing. That is, what if we gave it knowledge about multiple-agent situations (perhaps a referent group evaluation function for actions), knowledge about real-time interactions (perhaps rules about how to distinguish real-time from non-real time interactions and how to alter one's response in the two cases), knowledge about multiple goals (perhaps giving it individual level

goals such as hunger and thirst, group level goals such as maintenance of group identity and size, and social goals such as identifying with the organization it works for) knowledge about social structure (perhaps by giving it information that it has a particular age, sex, religion, job and the number of others in the same or different positions) and cultural-historical knowledge (perhaps by giving it knowledge about the norms, beliefs, expected actions, and expected response in a set of situations given the various social positions). The result would be an agent that would face new situations with equanimity, impartially observing and evaluating the situation, applying its knowledge and determining a course of action, an agent who, when faced with the same situation, would respond identically regardless of its health or well being, and an agent who when attacked would respond as its socio-culture dictated but without feeling an almost uncontrollable impulse to flee, and an agent who when given a classification or categorization task would respond only on the basis of this social knowledge. A collection of such agents, would seem to have little inventive capabilities and would seem to quickly come through interaction to act in a completely ritualized fashion. Moreover, there would be no room for individual differences other than those dictated by the socio-cultural situation. Such an agent, however "social" in knowledge would certainly seem to be devoid of the humanism one normally attributes to the true social agent. Such humanism, in our framework derives in large measure from the increased limitations present in an agent with emotional cognitive capabilities.[17]

Now, admittedly, giving such knowledge to Soar is no trivial task; nor is it necessarily straightforward as it involves situation description and enumeration of appropriate rules. Altering the cognitive agent into an emotional cognitive agent is, however, an even more difficult task (at least from the design standpoint). Preliminary investigation has led us to conclude that such transformation requires the agent to have motor sensory capabilities in addition to cognitive abilities.

The value of this enterprise, of this paper, lies in part in the utilization of this framework to determine how far we have come and where to go next in the development of an adequate theory of the social agent. To this end, we have introduced Soar, a model of the human cognitive agent and have noted where and hwo Soar must be augmented to construct a social agent. Additionally, part of the value of this enterprise lies in the proposed framework for laying out the nature of the social agent. This framework makes it possible not only to contrast, and hence locate the strong and weak points, of seemingly disparate theoretical approaches, but also to determine wherein lies the socialness. As the foregoing discussion exposed, social theories are for the most part pretty nonsocial. They gain most of their leverage not from social but from cognitive assumptions. Finally, part of the value of this enterprise come from a new perspective on the inherent socialness of humans. Socialness arises not from capabilities but from limitations. Socialness is a response to environmental complexity (the presence of multiple others, multiple and simultaneous goals, rich cultural-historical heritage, and so on). Our hope is that this perspective, of socialness as dependent on limited capabilities in a complex environment,

[17]We have used "seem to" throughout, because the whole point of the fractionation we propose is that we need to actually find out what is the case.

and the implied framework, detailing the nature of the limitations and complexities, will engender a more complete understanding of the social nature of human beings.

## REFERENCES


Ax, A. (1953) The physiological differentiation between fear and anger in humans. *Psychosomatic Medicine* **5**: 433–442.

Blau, P. M. (1977) *Inequality and Heterogeneity*, New York, NY: The Free Press of Macmillan Co.

Card, S., Moran, T. P., and Newell, A. (1983) *The Psychology of Human-Computer Interaction*, Hillsdale, NJ: Erlbaum.

Carley, K. M. (1990) Group stability: A socio-cognitive approach, in Lawler, E., Markovsky, B., Ridgeway, C., and Walker, H. (eds.) *Advances in Group Processes*, Greenwich, CN: JAI Press.

Carley, K. M. (1991) A theory of group stability. *American Sociological Review* **56**: 331–354.

Carley, K., Kjaer-Hansen, J., Prietula, M., and Newell, A. (1992) Plural-Soar: A prolegomenon to artificial agents and organizational behavior, in Masuch, M., and Warglien, M. (eds.) *Distributed Intelligence: Applications in human organizations*, Amsterdam, The Netherlands: Elsevier Science Publishers.

Eysenck, H. J., and Eysenck, M. W. (1985) *Personality and Individual Differences*, London: Elsevier.

Festinger, L. (1954) A theory of social comparison processes. *Human Relations* **7**: 117–140.

Frijda, N. (1987) *The Emotions*, New York, NY: Cambridge University Press.

Heise, D. (1977) Social action as the control of affect. *Behavioral Science* **22**: 163–177.

Heise, D. (1978) *Computer-Assisted Analysis of Social Action* (Tech. Rep.), Chapel Hill, NC: Institute for Research in Social Science.

Heise, D. (1979) *Understanding Events: Affect and the Construction of Social Action*, New York, NY: Cambridge University Press.

Huizinga, J. (1964) *Homo Ludens*, Boston, MA: Beacon.

Izard, C. (1972) *Patterns of Emotions: A New Analysis of Anxiety and Depression*, New York, NY: Academic Press.

Izard, C. (1977) *Human Emotions*, New York, NY: Plenum.

Kahneman, D., Slovic, P., and Tversky, A. (Eds.) (1982) *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge, England: Cambridge University Press.

Kemper, T. (1987) How many emotions are there? Wedding the social and the autonomic components. *American Journal of Sociology* **93**: 263–289.

Kogan, N., and Wallach, M. A. (1964) *Risk-Taking: A Study in Cognition and Personality*, New York: Holt, Rinhart and Winston.

Laird, J., Newell, A., and Rosenbloom, P. (1987) Soar: An architecture for general intelligence. *Artificial Intelligence* **33**: 1–64.

Laird, J., Rosenbloom, P., and Newell, A. (1986) *Universal Subgoaling and Chunking*, Boston, MA: Kluwer Academic Publisher.

Laird, J., Rosenbloom, P., and Newell, A. (1986) Chunking in Soar: The anatomy of a general learning mechanism. *Machine Learning* **1**: 11–46.

Lewis, R. L., Newell, A., and Polk, T. A. (1989) Toward a Soar theory of taking instructions for immediate reasoning tasks. *Proceedings Cognitive Science Eleventh Annual Conference, 1989*, Cognitive Science Society.

March, J. G., and Simon, H. A. (1958) *Organizations*, New York, NY: Wiley.

Marcuse (1964) *One-Dimensional Man: Studies int he Ideology of Advanced Industrial Society*, Boston, MA: Beacon Press.

Mead, G. H. (1934, 1962) *Mind, Self, and Society*, Chicago, IL: University of Chicago Press.

Newell, A. (1982) The knowledge level. *Artificial Intelligence* **18** (1): 87–127.

Newell, A. (1990) *Unified Theories of Cognition*, Cambridge, MA: Harvard University Press.

Norman, D. A. (1981) Twelve issues for cognitive science, in Norman, D. A. (ed.) *Perspectives on Cognitive Science*, Norwood, NJ: Ablex.

Ortony, A., Clore, G. L., and Collins, A. (1988) *Principia Pathematica: The Cognitive Structure of Emotions*, New York, NY: Cambridge University Press.

Ortony, A., Clore, G. L., and Foss, M. A. (1987) The referential structure of the affective lexicon. *Cognitive Science* **11**: 341–364.

Pitz, G. F., and Sachs, N. J. (1984) Judgment and decisions: Theory and application. *Annual Review of Psychology* **35**: 139–163.

Polk, T. A., and Newell, A. (1988) Modeling human syllogistic reasoning in Soar. *Proceedings Cognitive Science Tenth Annual Conference, 1988*. Cognitive Science Society.
Polk, T. A., Newell, A., and Lewis, R. L. (1989) Toward a unified theory of immediate reasoning in Soar. *Proceedings Cognitive Science Eleventh Annual Conference, 1989*, Cognitive Science Society.
Rapoport, A. (1961) *Fights, Games, and Debates*, Ann Arbor, MI: University of Michigan Press.
Ridgeway, D., Waters, E., and Kuczaj II, S. (1985) Acquisition of emotion-descriptive language: receptive and productive vocabulary norms for ages 18 months to 6 years. *Developmental Psychology* **21**: 901–908.
Ruiz, D., and Newell, A. (1989) Tower-noticing triggers strategy change in the Tower of Hanoi: A Soar model. *Proceedings Cognitive Science Eleventh Conference, 1989*, Cognitive Science Society.
Simon, H. A. (1957) *Administrative Behavior*, New York, NY: Macmillan.
Simon, H. A. (1976) From substantive to procedural rationality, in Latis, S. J. (ed.), *Method and Appraisal in Economics*, Cambridge: Cambridge University Press.
Simon, H. A. (1979) *Models of Thought*, New Haven, CT: Yale University Press.
Simon, H. (1983) *Models of Bounded Rationality*, Cambridge, MA: MIT Press.
Suls, J. M., and Miller, R. L. (eds.) (1977) *Social Comparison Processes: Theoretical and Empirical Perspectives*, New York: Washington Hemisphere Pub.
Turing, A. M. (1950) Computing machinery and intelligence. *Mind* **59**: 433–460. (Reprinted in Feigenbaum, E. A., and Feldman, J. (eds.) *Computers and Thought*, New York: McGraw-Hill, 1963).
Turner, J. (1988) *A Theory of Social Interaction*, Stanford, CA: Stanford University Press.
Tversky and Kahneman (1979) Prospect theory: An analysis of decision under risk. *Econometrica* **47** (2): 263–291.
Waldrop, M. M. (1989) Toward a unified theory of cognition. *Science* **241**: July 1.
Waldrop, M. M. (1989) Soar: A unified theory of cognition. *Science* **241**: July 15.
Wittich, C., and Roth, G. (eds.) Weber, M. (1978) *Economy and Society: An Outline of Interpretive Sociology*, Berkeley, CA.
Wuthnow, R. J. (1988) Sociology of Religion, in Smelser, N. J. (eds.) *Handbook of Sociology*, Beverly Hills, CA: Sage.